2018

First-order methods of solving nonconvex optimization problems: Algorithms, convergence, and optimality

Songtao Lu
Iowa State University

First-order methods of solving nonconvex optimization problems: Algorithms, convergence, and optimality

by

Songtao Lu

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Electrical Engineering

Program of Study Committee:
Mingyi Hong, Co-major Professor
Zhengdao Wang, Co-major Professor
Nicola Elia
Aleksandar Dogandžić
Kris De Brabanter

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2018

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

ABSTRACT

First-order methods for solving large scale nonconvex problems have been applied in many areas of machine learning, such as matrix factorization, dictionary learning, matrix sensing/completion, training deep neural networks, etc. For example, matrix factorization problems have lots of important applications in document clustering, community detection and image segmentation. In this dissertation, we first study some novel nonconvex variable splitting methods for solving some matrix factorization problems, mainly focusing on symmetric non-negative matrix factorization (SymNMF) and stochastic SymNMF.

In the problem of SymNMF, the proposed algorithm, called nonconvex splitting SymNMF (NS-SymNMF), is guaranteed to converge to the set of Karush-Kuhn-Tucker (KKT) points of the nonconvex SymNMF problem. Furthermore, it achieves a global sublinear convergence rate. We also show that the algorithm can be efficiently implemented in a distributed manner. Further, sufficient conditions are provided which guarantee the global and local optimality of the obtained solutions. Extensive numerical results performed on both synthetic and real data sets suggest that the proposed algorithm converges quickly to a local minimum solution.

Furthermore, we consider a stochastic SymNMF problem in which the observation matrix is generated in a random and sequential manner. The proposed stochastic nonconvex splitting method not only guarantees convergence to the set of stationary points of the problem (in the mean-square sense), but further achieves a sublinear convergence rate. Numerical results show that for clustering problems over both synthetic and real world datasets, the proposed algorithm converges quickly to the set of stationary points.

When the objective function is nonconvex, it is well-known the most of the first-order algorithms converge to the first-order stationary solution (SS1) with a global sublinear rate. Whether the first-order algorithm can converge to the second-order stationary points (SS2) with some provable rate

has attracted a lot of attention recently. In particular, we study the alternating gradient descent (AGD) algorithm as an example, which is a simple but popular algorithm and has been applied to problems in optimization, machine learning, data mining, and signal processing, etc. The algorithm updates two blocks of variables in an alternating manner, in which a gradient step is taken on one block, while keeping the remaining block fixed.

In this work, we show that a variant of AGD-type algorithms will not be trapped by "bad" stationary solutions such as saddle points and local maximum points. In particular, we consider a smooth unconstrained nonconvex optimization problem, and propose a perturbed AGD (PA-GD) which converges (with high probability) to the set of SS2 with a global sublinear rate. To the best of our knowledge, this is the first alternating type algorithm which is guaranteed to achieve SS2 points with high probability and the corresponding convergence rate is $\mathcal{O}(\text{polylog}(d)/\epsilon^{7/3})$ [where $\text{polylog}(d)$ is polynomial of the logarithm of problem dimension d].

CHAPTER 1. OVERVIEW

In this dissertation, we study nonconvex optimization problems in both constrained and unconstrained cases. For the constrained case, we consider the symmetric nonnegative matrix factorization (SymNMF) as an example. We propose a nonconvex splitting algorithm and study the convergence behaviour of the algorithm. Also, the optimality condition of the obtained solutions for this problem is provided, which verifies the quality of the obtained solutions. Furthermore, the stochastic SymNMF is also considered, where the corresponding stochastic nonconvex splitting algorithm is proposed as well. In the unconstrained nonconvex optimization case, there are many nonconvex optimization problems where the saddle points of the objective functions are strict and local optimal points are also global ones. Then, a perturbed alternating gradient descent algorithm is proposed for solving a class of block structured nonconvex optimization problems. Finally, under some mild assumptions, we will show that the proposed algorithm is able to converge to the second-order stationary points (SS2) with high probability in a sublinear rate.

1.1 Constrained Nonconvex Problems

1.1.1 Symmetric Nonnegative Matrix Factorization

Non-negative matrix factorization (NMF) refers to factoring a given matrix into the product of two matrices whose entries are all non-negative. It has long been recognized as an important matrix decomposition problem (1; 2). The requirement that the factors are component-wise nonnegative makes the NMF distinct from traditional methods such as the principal component analysis (PCA) and latent dirichlet allocation (LDA), leading to many interesting applications in imaging, signal processing and machine learning (3; 4; 5; 6; 7); see (8) for a recent survey. When further requiring that the two factors are identical after transposition, the NMF becomes the so-called SymNMF. In the case where the given matrix cannot be factorized exactly, an approximate solution with

a suitably defined approximation error is desired. Mathematically, the SymNMF approximates a given (usually symmetric) non-negative matrix $\mathbf{Z} \in \mathbb{R}^{N\times N}$ by a low rank matrix $\mathbf{X}\mathbf{X}^T$, where the factor matrix $\mathbf{X} \in \mathbb{R}^{N\times K}$ is component-wise non-negative, typically with $K \ll N$. Such problem can be formulated as the following nonconvex optimization problem (9; 10; 11):

$$\min_{\mathbf{X}\geq 0} \quad \frac{1}{2}\|\mathbf{X}\mathbf{X}^T - \mathbf{Z}\|_F^2 \tag{1.1}$$

where $\|\cdot\|_F$ denotes the Frobenius norm and inequality constraint $\mathbf{X} \geq 0$ is component-wise.

Recently, SymNMF has found many applications in document clustering, community detection, image segmentation and pattern clustering in bioinformatics (11; 12; 13; 9). An important class of clustering methods is known as the spectral clustering, e.g., (14; 15), which is based on the eigenvalue decomposition of some transformed graph Laplacian matrix. In (16), it has been shown that spectral clustering and SymNMF are two different ways of relaxing the kernel K-means clustering, where the former relaxes the nonnegativity constraint while the latter relaxes certain orthogonality constraint. Furthermore, SymNMF has the advantage that it often yields more meaningful and interpretable results (11). In this work, a new nonconvex splitting method is proposed which is an efficient way of solving SymNMF problems with provable convergence guarantees.

1.1.2 Stochastic SymNMF

Classical SymNMF problems in the data mining area are deterministic, where the observation matrix $\mathbf{Z}$ is completely known (11; 17). However, in recent applications such as social network community detection, the matrix $\mathbf{Z}$ represents the relations among the clusters/communities, observed during a given time period. By nature such matrix is random, whose structure is determined by the dynamics of the network connections (18). Furthermore, in many modern big-data related problems such as matrix completion (19), subspace tracking (20), community detection, the data are usually collected through some random sampling techniques. As a concrete example, in community detection problems the observed activities among the nodes can change over time hence is random. In these applications sampling the connectivity of the graph at a given time results in a random similarity matrix, such as stochastic block model (21). Mathematically, the stochastic

SymNMF problem can be formulated as the following stochastic optimization problem

$$\min_{\mathbf{X}\geq 0} \quad \frac{1}{2}\mathbb{E}_{\mathbf{Z}}[\|\mathbf{X}\mathbf{X}^T - \mathbf{Z}\|_F^2] \tag{1.2}$$

where $\mathbf{Z}$ follows some distribution over a set $\Xi \in \mathbb{R}^{N\times N}$, and the expectation is taken over the random observation $\mathbf{Z}$. In clustering problems, the samples of matrix $\mathbf{Z}$ can be the similarity matrix which measures the connections among nodes over networks.

As we will see later, the problem in (1.1) is equivalent to $\min_{\mathbf{X}\geq 0} \|\mathbf{X}\mathbf{X}^T - \mathbb{E}_{\mathbf{Z}}[\mathbf{Z}]\|_F^2$. If we know the distribution of $\mathbf{Z}$, then we can computer $\mathbb{E}_{\mathbf{Z}}[\mathbf{Z}]$ first and the problem is converted to a classical SymNMF problem. However, in practice, we usually do not have access to the underlying distribution of $\mathbf{Z}$. Instead, we can obtain sequentially realizations of $\mathbf{Z}$, such as in the application of online streaming data (22). It is possible to use a batch of samples to compute the empirical mean of $\mathbf{Z}$ and implement the deterministic SymNMF algorithms. As more samples are collected, the empirical mean will converge to the ensemble mean, leading to a consistent estimator of the solution of the symmetric factor $\mathbf{X}$. There are two problems with such an approach. First, it may be desirable to have an estimate of the symmetric factor $\mathbf{X}$ at each time instant, namely when a new sample of $\mathbf{Z}$ is available. Running the complete SymNMF algorithm at each time instant may be computationally expensive. Second, even if the computational complexity is not a concern, existing analysis results and theoretical guarantees such as convergence rate are not applicable to the case where the matrix to be factorized is changing with time (although eventually converging to the ensemble mean). Therefore, it is desirable to develop efficient algorithms that produce online SymNMF updates based on sequential realizations of $\mathbf{Z}$.

1.2 Unconstrained Nonconvex Problems

Although the constrained nonconvex problems have been solved efficiently, these first-order can only guraantee that the generated squence by the algorithms converge to the first-order stationary points (SS1). In recent works, it has been shown that with some new techniques, such as adding some perturbation on the iterates of the algorithm occasionally, the first-order algorithms can converge to second-order stationary points (SS2) efficiently. In this work, we take one of the most

popular algorithm, alternating gradient descent (AGD), as example and study the convergence behaviour of this algorithm to SS2.

1.2.1 Perturbed Alternating Gradient Descent

We consider a smooth and unconstrained nonconvex optimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^{d\times 1}} f(\mathbf{x}) \tag{1.3}$$

where $f:\mathbb{R}^d\to\mathbb{R}$ is twice differentiable. Problem (1.3) is a general formulation in most of machine learning topics, such as matrix factorization-type of problems (23; 24), regression problems (25), deep learning problems (26).

There are many ways of solving problem (1.3), such as gradient descent (GD), accelerated gradient descent (AGD), etc. When the problem dimension is large, it is natural to split the variables into multiple blocks and solve the subproblems with smaller size individually. The block coordinate descent (BCD) algorithm, and many of its variants such as block coordinate gradient descent (BCGD) and alternating gradient descent (AGD) (27; 28), are among the most powerful tools for solving large scale convex/nonconvex optimization problems (29; 30; 31; 32; 33). The BCD-type algorithms partition the optimization variables into multiple small blocks, and optimize each block one by one following certain block selection rule, such as cyclic rule (34), Gauss-Southwell rule (35), etc. To be more specific, problem (1.3) can be solved by the following reformulation.

$$\min_{\mathbf{x}_k} f(\mathbf{x}_1,\ldots,\mathbf{x}_k,\ldots,\mathbf{x}_K),\quad k=1,\ldots,K \tag{1.4}$$

where k denotes the index of the blocks, and K denotes the total number of blocks.

In recent years, there are many applications of BCD-type algorithms in the areas of machine learning and data mining, such as matrix factorization (36), tensor decomposition, low rank matrix estimation (37; 23), matrix completion/sensing (19), and training deep neural networks (DNNs) (38). Under relatively mild conditions, the convergence of BCD-type algorithms to SS1 have been broadly investigated for nonconvex and non-differentiable optimization (34; 39; 40). In particular, it is known that under mild conditions, these algorithms also achieve global sublinear rates (41).

However, despite its popularity and significant recent progress in understanding its behavior, it remains unclear whether BCD-type algorithms can converge to the set of SS2 with a provable global rate, even for the simplest problem with two blocks of variables.

Algorithms that can escape from strict saddle points – those stationary points that have negative eigenvalues – have wide applications. Many recent works have analyzed the saddle points in machine learning problems (42). Such as learning in shallow networks, the stationary points are either global minimum points or strict saddle points. In two-layer porcupine neural networks (PNNs), it has been shown that most local optima of PNN optimizations are also global optimizers (43). Previous work in (44) has shown that the saddle points in tensor decomposition are indeed strict saddle points. Also, it has been shown that any saddle points are strict in dictionary learning and phase retrieval problems theoretically (45; 46) and numerically in (47). More recently, (24) proposed a unified analysis of saddle points for a board class of low rank matrix factorization problems, and they proved that these saddle points are strict.

Motivated by these results, we will show that AGD with some random perturbation can still converge to SS2 efficiently for unconstrained nonconvex optimization problems in a global sublinear convergence rate.

CHAPTER 2. SYMMETRIC NONNEGATIVE MATRIX FACTORIZATION

2.1 Introduction

Due to the importance of the NMF problem, many algorithms have been proposed in the literature for finding its high-quality solutions. Well-known algorithms include the multiplicative update (6), alternating projected gradient methods (48), alternating nonnegative least squares (ANLS) with the active set method (49) and a few recent methods such as the bilinear generalized approximate message passing (50; 51), as well as methods based on the block coordinate descent (52). These methods often possess strong convergence guarantees (to Karush-Kuhn-Tucker (KKT) points of the NMF problem) and most of them lead to satisfactory performance in practice; see (8) and the references therein for detailed comparison and comments for different algorithms. Unfortunately, most of the aforementioned methods for NMF lack effective mechanisms to enforce the symmetry between the resulting factors, therefore they are not directly applicable to the SymNMF. Recently, there have been a number of works that focus on designing customized algorithms for SymNMF, which we review below.

2.1.1 Related Work

To this end, first rewrite the SymNMF equivalently as

$$\min_{\mathbf{Y}\geq 0, \mathbf{X}=\mathbf{Y}} \frac{1}{2}\|\mathbf{X}\mathbf{Y}^T - \mathbf{Z}\|_F^2. \tag{2.1}$$

A simple strategy is to ignore the equality constraint $\mathbf{X} = \mathbf{Y}$, and then alternatingly perform the following two steps: 1) solving $\mathbf{Y}$ with $\mathbf{X}$ being fixed (a non-negative least squares problem); 2) solving $\mathbf{X}$ with $\mathbf{Y}$ being fixed (a least squares problem). Such ANLS algorithm has been proposed in (11) for dealing with SymNMF. Unfortunately, despite the fact that an optimal solution can be obtained in each subproblem, there is no guarantee that the $\mathbf{Y}$-iterate will converge to the $\mathbf{X}$-iterate. The algorithm in (11) adds a regularized term for the difference between the two factors to

the objective function and explicitly enforces that the two matrices be equal at the output. Such an extra step enforces symmetry, but unfortunately also leads to the loss of global convergence guarantee. A related ANLS-based method has been introduced in (10); however the algorithm is based on the assumption that there exists an exact symmetric factorization (i.e., $\exists\ \mathbf{X} \geq 0$ such $\mathbf{X}\mathbf{X}^T = \mathbf{Z}$). Without such assumption, the algorithm may not converge to the set of KKT points[1] of (1.1). A multiplicative update for SymNMF has been proposed in (9), but the algorithm lacks convergence guarantee (to KKT points of (1.1)) (53), and has a much slower convergence speed than the one proposed in (10). In (11; 54), algorithms based on the projected gradient descent (PGD) and the projected Newton (PNewton) have been proposed, both of which directly solve the original formulation (1.1). Again there has been no global convergence analysis since the objective function is a nonconvex fourth-order polynomial. More recently, the work (55) applies the nonconvex coordinate descent (CD) algorithm for SymNMF. However, due to the fact that the minimizer of the fourth order polynomial is not unique in each coordinate updating, the CD-based method may not converge to stationary points.

Another popular method for NMF is based on the alternating direction method of multipliers (ADMM), which is a flexible tool for large scale convex optimization (56). For example, using ADMM for both NMF and matrix completion, high quality results have been obtained in (57) for gray-scale and hyperspectral image recovery. Furthermore, ADMM has been applied to generalized versions of NMF where the objective function is the general beta-divergence (58). A hybrid alternating optimization and ADMM method was proposed for NMF, as well as tensor factorization, under a variety of constraints and loss measures in (59). However, despite the promising numerical results, none of the works discussed above has rigorous theoretical justification for SymNMF. Technically, imposing symmetry poses much difficulty in the analysis (we will comment on this point shortly). In fact, the convergence of ADMM for SymNMF is still open in the literature.

An important research question for NMF and SymNMF is whether it is possible to design algorithms that lead to *globally* optimal solutions. At the first sight such problem appears very

[1] Let $d(a, s)$ denote the distance between two points a and s. We say that a sequence a_i converges to a set $\mathcal{S}$ if the distance between a_i and $\mathcal{S}$, defined as $\inf_{s\in\mathcal{S}} d(a_i, s)$, converges to zero, as $i \to \infty$.

challenging since finding the exact NMF is NP-hard (60) and checking whether a positive semidefinite matrix can be decomposed exactly by SymNMF is also NP-hard (61). However, some promising recent findings suggest that when the structure of the underlying factors are appropriately utilized, it is possible to obtain rather strong results. For example, in (62), the authors have shown that for the low rank factorized stochastic optimization problem where the two low rank matrices are symmetric, a modified stochastic gradient descent algorithm is capable of converging to a global optimum with constant probability from a random starting point. Related works also include (63; 64; 36). However, when the factors are required to be non-negative and symmetric, it is no longer clear whether the existing analysis can still be used to show convergence to global optimal points, even local optimality (a milder result). For the non-negative principal component problem (that is, finding the leading non-negative eigenvector, i.e., $K = 1$), under the spiked model, reference (65) shows that certain approximate message passing algorithm is able to find the global optimal solution asymptotically. Unfortunately, this analysis does not generalize to an arbitrary symmetric observation matrix with a larger K. To our best knowledge, there is a lack of characterization of global and local optimal solutions for the SymNMF problem.

2.1.2 Contributions

In this work, we first propose a novel algorithm for the SymNMF, which utilizes nonconvex splitting and is capable of converging to the set of KKT points with provable global convergence rate. The main idea is to relax the symmetry requirement at the beginning and gradually enforce it as the algorithm proceeds. Second, we provide a number of easy-to-check sufficient conditions guaranteeing the local or global optimality of the obtained solutions. Numerical results on both synthetic and real data show that the proposed algorithm achieves fast and stable convergence (often to local minimum solutions) with low computational complexity.

More specifically, the main contributions of this work are:

1) We design a novel algorithm, named the nonconvex splitting SymNMF (NS-SymNMF), which converges to the set of KKT points of SymNMF with a global sublinear rate. To our best knowledge, it is the first SymNMF solver that possesses global convergence rate guarantee.

2) We provide a set of easily checkable sufficient conditions (which only involve finding the smallest eigenvalue of certain matrix) that characterize the global and local optimality of the SymNMF. By utilizing such conditions, we demonstrate numerically that with high probability, our proposed algorithm converges not only to the set of KKT points but to a local optimal solution as well.

Notation: Bold upper case letters without subscripts (e.g., $\mathbf{X}, \mathbf{Y}$) denote matrices and bold lower case letters without subscripts (e.g., $\mathbf{x}, \mathbf{y}$) represent vectors. The notation $\mathbf{Z}_{i,j}$ denotes the (i,j)-th entry of the matrix $\mathbf{Z}$. The vector $\mathbf{X}_i$ denotes the ith row of the matrix $\mathbf{X}$ and $\mathbf{X}'_m$ denotes the mth column of the matrix. The letter $\mathcal{Y}$ denotes the feasible set of an optimization variable $\mathbf{Y}$.

2.2 NS-SymNMF

The proposed algorithm leverages the reformulation (2.1). Our main idea is to gradually tighten the difficult equality constraint $\mathbf{X} = \mathbf{Y}$ as the algorithm proceeds so that when convergence is approached, such equality is eventually satisfied. To this end, let us construct the augmented Lagrangian for (2.1), given by

$$\mathcal{L}(\mathbf{X}, \mathbf{Y}; \mathbf{\Lambda}) = \frac{1}{2}\|\mathbf{X}\mathbf{Y}^T - \mathbf{Z}\|_F^2 + \langle \mathbf{Y} - \mathbf{X}, \mathbf{\Lambda} \rangle + \frac{\rho}{2}\|\mathbf{Y} - \mathbf{X}\|_F^2 \tag{2.2}$$

where $\mathbf{\Lambda} \in \mathbb{R}^{N \times K}$ is a matrix of dual variables, $\langle \cdot \rangle$ denotes the inner product operator, and $\rho > 0$ is a penalty parameter whose value will be determined later.

At this point, it may be tempting to directly apply the well-known ADMM method to the augmented Lagrangian (2.2), which alternatingly minimizes the primal variables $\mathbf{X}$, $\mathbf{Y}$, followed by a dual ascent step $\mathbf{\Lambda} \leftarrow \mathbf{\Lambda} + \rho(\mathbf{Y} - \mathbf{X})$. Unfortunately, the classical result for ADMM presented in (56; 66; 67) only works for convex problems, hence they do not apply to our nonconvex problem (2.1) (note this is a linearly constrained *nonconvex* problem where the nonconvexity arises in the

objective function). Recent results such as (68; 69; 70; 71) that analyze ADMM for nonconvex problems do not apply either, because in these works the basic requirements are: 1) the objective function is separable over the block variables; 2) the smooth part of the augmented Lagrangian function has Lipschitz continuous gradient with respect to all variable blocks. Unfortunately neither of these conditions are satisfied in our problem.

Next we begin presenting the proposed algorithm. We start by considering the following reformulation of problem (1.1)

$$\begin{aligned} \min_{\mathbf{X},\mathbf{Y}} \quad & \frac{1}{2}\|\mathbf{X}\mathbf{Y}^T - \mathbf{Z}\|_F^2 \\ \text{s.t.} \quad & \mathbf{Y} \geq 0,\ \mathbf{X} = \mathbf{Y},\ \|\mathbf{Y}_i\|_2^2 \leq \tau,\ \forall\ i, \end{aligned} \tag{2.3}$$

where $\mathbf{Y}_i$ denotes the ith row of the matrix $\mathbf{Y}$; $\tau > 0$ is some given constant. It is easy to check that when τ is sufficiently large (with a lower bound dependent on $\mathbf{Z}$), then problem (2.3) is *equivalent* to problem (1.1), implying that the KKT points $\mathbf{X}^*$ of the two problems are identical, where the KKT conditions of problem (1.1) are given by (72, eq. (5.49))

$$2\left(\mathbf{X}^*(\mathbf{X}^*)^T - \frac{\mathbf{Z}^T + \mathbf{Z}}{2}\right)\mathbf{X}^* - \mathbf{\Omega}^* = 0, \tag{2.4a}$$

$$\mathbf{\Omega}^* \geq 0, \tag{2.4b}$$

$$\mathbf{X}^* \geq 0, \tag{2.4c}$$

$$\mathbf{X}^* \circ \mathbf{\Omega}^* = \mathbf{0} \tag{2.4d}$$

where $\mathbf{\Omega}^*$ is the dual matrix for the constraint $\mathbf{X} \geq 0$ and $\circ$ denotes the Hadamard product.

Also, the points $\mathbf{X}^*$ are the KKT points of the SymNMF problem if and only if they are the stationary points of SymNMF which satisfy the optimality conditions given by (27, Proposition 2.1.2)

$$\left\langle \left(\mathbf{X}^*(\mathbf{X}^*)^T - \frac{\mathbf{Z}^T + \mathbf{Z}}{2}\right)\mathbf{X}^*, \mathbf{X} - \mathbf{X}^* \right\rangle \geq 0, \quad \forall\ \mathbf{X} \geq 0. \tag{2.5}$$

To be precise, we have the following results.

Lemma 1. *The KKT points and stationary points of the SymNMF problem are equivalent.*

Proof: See Section A.1

Lemma 2. *Suppose* $\tau > \theta_k, \forall k$ *where*

$$\theta_k \triangleq \frac{\mathbf{Z}_{k,k} + \frac{1}{2}\sqrt{\sum_{i=1}^{N}(\mathbf{Z}_{i,k} + \mathbf{Z}_{k,i})^2}}{2}, \tag{2.6}$$

then the KKT points of the problem (1.1) *and those of* (2.3) *have a one-to-one correspondence.*

Proof: See Section A.2.

We remark that the previous work (55) has made the observation that solving SymNMF with the additional constraints $\|\mathbf{X}_i\|_2 \leq \sqrt{2\|\mathbf{Z}\|_F}, \forall i$ will not result in any loss of the *global* optimality. Lemma 2 provides a stronger result, that all *KKT* points of SymNMF are preserved within a *smaller* bounded feasible set $\mathcal{Y} \triangleq \{\mathbf{Y} \mid \mathbf{Y}_i \geq 0, \|\mathbf{Y}_i\|_2^2 \leq \tau, \forall i\}$ (note, that $\tau \ll 2\|\mathbf{Z}\|_F$ in general).

The proposed algorithm, named the nonconvex splitting SymNMF (NS-SymNMF), alternates between the primal updates of variables $\mathbf{X}$ and $\mathbf{Y}$, and the dual update for $\mathbf{\Lambda}$. Below we present its detailed steps (superscript t is used to denote the iteration number).

$$\mathbf{Y}^{(t+1)} = \arg\min_{\mathbf{Y}\geq 0, \|\mathbf{Y}_i\|_2^2 \leq \tau, \forall i} \frac{1}{2}\|\mathbf{X}^{(t)}\mathbf{Y}^T - \mathbf{Z}\|_F^2 + \frac{\rho}{2}\|\mathbf{Y} - \mathbf{X}^{(t)} + \mathbf{\Lambda}^{(t)}/\rho\|_F^2 + \frac{\beta^{(t)}}{2}\|\mathbf{Y} - \mathbf{Y}^{(t)}\|_F^2, \tag{2.7}$$

$$\mathbf{X}^{(t+1)} = \arg\min_{\mathbf{X}} \frac{1}{2}\|\mathbf{X}(\mathbf{Y}^{(t+1)})^T - \mathbf{Z}\|_F^2 + \frac{\rho}{2}\|\mathbf{X} - \mathbf{\Lambda}^{(t)}/\rho - \mathbf{Y}^{(t+1)}\|_F^2, \tag{2.8}$$

$$\mathbf{\Lambda}^{(t+1)} = \mathbf{\Lambda}^{(t)} + \rho(\mathbf{Y}^{(t+1)} - \mathbf{X}^{(t+1)}), \tag{2.9}$$

$$\beta^{(t+1)} = \frac{6}{\rho}\|\mathbf{X}^{(t+1)}(\mathbf{Y}^{(t+1)})^T - \mathbf{Z}\|_F^2. \tag{2.10}$$

We remark that this algorithm is very close in form to the standard ADMM method applied to problem (2.3) (which lacks convergence guarantees). The key difference is the use of the proximal term $\|\mathbf{Y} - \mathbf{Y}^{(t)}\|_F^2$ multiplied by an *iteration dependent* penalty parameter $\beta^{(t)} \geq 0$, whose value is proportional to the size of the objective value. Intuitively, if the algorithm converges to a solution with small objective value (which appears to be often the case in practice based on our numerical experiments), then the parameter $\beta^{(t)}$ vanishes in the limit. Introducing such proximal term is one of the main novelty of the algorithm, and it is crucial in guaranteeing the convergence of NS-SymNMF.

2.3 Convergence Analysis

In this section we provide convergence analysis of the NS-SymNMF for a general SymNMF problem. We do not require $\mathbf{Z}$ to be symmetric, positive-semidefinite, or to have positive entries. We assume K can take any integer value in $[1,\ N]$.

2.3.1 Convergence and Convergence Rate

Below we present our first main result, which asserts that when the penalty parameter ρ is sufficiently large, the NS-SymNMF algorithm converges globally to the set of KKT points of (1.1).

Theorem 1. *Suppose the following is satisfied*

$$\rho > 6N\tau. \tag{2.11}$$

Then the following statements are true for NS-SymNMF:

1. *The equality constraint is satisfied in the limit, i.e.,*

$$\lim_{t\to\infty} \|\mathbf{X}^{(t)} - \mathbf{Y}^{(t)}\| \to 0.$$

2. *The sequence $\{\mathbf{X}^{(t)}, \mathbf{Y}^{(t)}; \mathbf{\Lambda}^{(t)}\}$ generated by the algorithm is bounded. And every limit point of the sequence is a KKT point of problem (1.1).*

An equivalent statement on the convergence is that the sequence $\{\mathbf{X}^{(t)}, \mathbf{Y}^{(t)}; \mathbf{\Lambda}^{(t)}\}$ converges to the set of KKT points of problem (1.1); cf. footnote 1 on Page 7.

Proof: See Section A.3.

Our second result characterizes the convergence rate of the algorithm. To this end, we need to construct a function that measures the optimality of the iterates $\{\mathbf{X}^{(t)}, \mathbf{Y}^{(t)}; \mathbf{\Lambda}^{(t)}\}$. Define the *proximal gradient* of the augmented Lagrangian function as

$$\widetilde{\nabla}\mathcal{L}(\mathbf{X}, \mathbf{Y}; \mathbf{\Lambda}) \triangleq \begin{bmatrix} \mathbf{Y}^T - \mathsf{proj}_{\mathcal{Y}}[\mathbf{Y}^T - \nabla_{\mathbf{Y}}(\mathcal{L}(\mathbf{Y}, \mathbf{X}; \mathbf{\Lambda})] \\ \nabla_{\mathbf{X}}\mathcal{L}(\mathbf{X}, \mathbf{Y}; \mathbf{\Lambda}) \end{bmatrix} \tag{2.12}$$

where the operator

$$\text{proj}_{\mathcal{Y}}(\mathbf{W}) \triangleq \arg \min_{\mathbf{Y}\geq 0, \|\mathbf{Y}_i\|_2^2 \leq \tau, \forall i} \|\mathbf{W}-\mathbf{Y}\|_F^2 \tag{2.13}$$

i.e., it is the projection operator that projects a given matrix $\mathbf{W}$ onto the feasible set of $\mathbf{Y}$. Here we propose to use the following quantity to measure the progress of the algorithm

$$\mathcal{P}(\mathbf{X}^{(t)}, \mathbf{Y}^{(t)}; \mathbf{\Lambda}^{(t)}) \triangleq \|\widetilde{\nabla}\mathcal{L}(\mathbf{X}^{(t)}, \mathbf{Y}^{(t)}; \mathbf{\Lambda}^{(t)})\|_F^2 + \|\mathbf{X}^{(t)} - \mathbf{Y}^{(t)}\|_F^2. \tag{2.14}$$

It can be verified that if $\lim_{t\to\infty} \mathcal{P}(\mathbf{X}^{(t)}, \mathbf{Y}^{(t)}; \mathbf{\Lambda}^{(t)}) = 0$, then a KKT point of problem (1.1) is obtained.

Below we show that the function $\mathcal{P}(\mathbf{X}^{(t)}, \mathbf{Y}^{(t)}; \mathbf{\Lambda}^{(t)})$ goes to zero in a sublinear manner.

Theorem 2. *For a given small constant ϵ, let $T(\epsilon)$ denote the iteration index satisfying the following inequality*

$$T(\epsilon) \triangleq \min\{t \mid \mathcal{P}(\mathbf{X}^{(t)}, \mathbf{Y}^{(t)}; \mathbf{\Lambda}^{(t)}) \leq \epsilon, t \geq 0\}. \tag{2.15}$$

Then there exists some constant $C > 0$ such that

$$\epsilon \leq \frac{C\mathcal{L}(\mathbf{X}^{(1)}, \mathbf{Y}^{(1)}; \mathbf{\Lambda}^{(1)})}{T(\epsilon)}. \tag{2.16}$$

Proof: See Section A.4. The above result indicates that in order for $\mathcal{P}(\mathbf{X}^{(t)}, \mathbf{Y}^{(t)}; \mathbf{\Lambda}^{(t)})$ to reach below ϵ, it takes $\mathcal{O}(1/\epsilon)$ number of iterations. It follows that NS-SymNMF converges sublinearly.

2.3.2 Sufficient Global and Local Optimality Conditions

Since problem (1.1) is not convex, the KKT points obtained by NS-SymNMF could be different from the global optimal solutions. Therefore it is important to characterize the conditions under which these two different types of solutions coincide. Below we provide an easily checkable sufficient condition to ensure that a KKT point $\mathbf{X}^*$ is also a globally optimal solution for problem (1.1).

Theorem 3. *Suppose that $\mathbf{X}^*$ is a KKT point of (1.1). Then, $\mathbf{X}^*$ is also a global optimal point if the following is satisfied*

$$\mathbf{S} \triangleq \mathbf{X}^*(\mathbf{X}^*)^T - \frac{\mathbf{Z}^T + \mathbf{Z}}{2} \succeq 0. \tag{2.17}$$

Proof: See Section A.5.

It is important to note that condition (2.17) is only a sufficient condition and hence may be difficult to satisfy in practice. In this section we provide a milder condition which ensures that a KKT point is *locally optimal*. This type of result is also very useful in practice since it can help identify spurious saddle points such as the point $\mathbf{X}^* = \mathbf{0}$ in the case where $\mathbf{Z}^T + \mathbf{Z}$ is not negative semidefinite.

We have the following characterization of the local optimal solution of the SymNMF problem.

Theorem 4. *Suppose that $\mathbf{X}^*$ is a KKT point of (1.1). Define a matrix $\mathcal{T} \in \mathbb{R}^{KN \times KN}$ whose (m,n)th block is a matrix of size $N \times N$*

$$\mathcal{T}_{m,n} \triangleq \left((\mathbf{X}_m'^*)^T \mathbf{X}_n'^* - \delta \|\mathbf{X}_n'^*\|_2^2\right) \mathbf{I} + \mathbf{X}_n'^* (\mathbf{X}_m'^*)^T + \delta_{m,n} \mathbf{S}, \tag{2.18}$$

where $\mathbf{S}$ is defined in (2.17), $\delta_{m,n}$ is the Kronecker delta function, and $\mathbf{X}_m'^*$ denotes the mth column of $\mathbf{X}^*$. If there exists some $\delta > 0$ such that $\mathcal{T} \succ 0$, then $\mathbf{X}^*$ is a strict local minimum solution of (1.1), meaning that there exists some $\epsilon > 0$ small enough such that for all $\mathbf{X} \geq 0$ satisfying $\|\mathbf{X} - \mathbf{X}^*\|_F \leq \epsilon$, we have

$$f(\mathbf{X}) \geq f(\mathbf{X}^*) + \frac{\gamma}{2} \|\mathbf{X} - \mathbf{X}^*\|_F^2. \tag{2.19}$$

Here the constant $\gamma > 0$ is given by

$$\gamma = -\left(\frac{2K^2}{\delta} + K(K-2)\right) \epsilon^2 + 2\lambda_{\min}(\mathcal{T}) > 0 \tag{2.20}$$

where $\lambda_{\min}(\mathcal{T}) > 0$ is the smallest eigenvalue of $\mathcal{T}$.

Proof: See Section A.6.

In the special case of $K = 1$, the sufficient condition set forth in Theorem 4 can be significantly simplified.

Corollary 1. *Suppose that $\mathbf{x}^*$ is the KKT point of (1.1) when $K = 1$. If there exists some $\delta > 0$ such that*

$$\mathcal{T}_1 \triangleq (1-\delta)\|\mathbf{x}^*\|_2^2 \mathbf{I} + 2\mathbf{x}^*(\mathbf{x}^*)^T - \frac{\mathbf{Z}^T + \mathbf{Z}}{2} \succ 0, \tag{2.21}$$

then $\mathbf{x}^*$ is a strict local minimum point of (1.1).

Proof: See Section A.7.

We comment that the condition given in Theorem 4 is much milder than that in Theorem 3. Further such condition is also very easy to check as it only involves finding the smallest eigenvalue of a $KN \times KN$ matrix for a given δ [2]. In our numerical result (to be presented shortly), we set a series of consecutive δ when performing the test. We have observed that the solutions generated by the proposed NS-SymNMF algorithm satisfy the condition provided in Theorem 4 with high probability.

2.3.3 Implementation

In this section we discuss the implementation of the proposed algorithm.

2.3.3.1 The X-Subproblem

The subproblem for updating $\mathbf{X}^{(t+1)}$ in (2.8) is equivalent to the following problem

$$\min_{\mathbf{X}} \|\mathbf{Z}_{\mathbf{X}}^{(t+1)} - \mathbf{X}\mathbf{A}_{\mathbf{X}}^{(t+1)}\|_F^2 \tag{2.22}$$

where

$$\begin{aligned} \mathbf{Z}_{\mathbf{X}}^{(t+1)} &\triangleq \mathbf{Z}\mathbf{Y}^{(t+1)} + \mathbf{\Lambda}^{(t)} + \rho\mathbf{Y}^{(t+1)} \\ \mathbf{A}_{\mathbf{X}}^{(t+1)} &\triangleq (\mathbf{Y}^{(t+1)})^T\mathbf{Y}^{(t+1)} + \rho\mathbf{I} \succ 0 \end{aligned} \tag{2.23}$$

are two fixed matrices. Clearly problem (2.22) is just a least-square problem and can be solved in closed-form. The solution is given by

$$\mathbf{X}^{(t+1)} = \mathbf{Z}_{\mathbf{X}}^{(t+1)}(\mathbf{A}_{\mathbf{X}}^{(t+1)})^{-1}. \tag{2.24}$$

We remark that the $\mathbf{A}_{\mathbf{X}}^{(t+1)}$ is a $K \times K$ matrix, where K is usually small (e.g., the number of clusters for graph clustering applications). As a result, $\mathbf{X}^{(t+1)}$ in (2.24) can be obtained by solving a small system of linear equations and hence computationally cheap.

[2] To find such smallest eigenvalue, we can find the largest eigenvalue of $\eta\mathbf{I} - \mathcal{T}$, using algorithms such as the power method (15), where η is sufficient large based on τ and $\|\mathbf{Z}\|_F$.

2.3.3.2 The Y-Subproblem

To solve the $\mathbf{Y}$-subproblem (2.7), we can use the gradient projection method. This problem can be decomposed into N separable constrained least squares problems, each of which can be solved independently, and hence can be implemented in parallel. Here we use the conventional gradient projection (GP) for solving each subproblem, which generates a sequence by

$$\mathbf{Y}_i^{(r+1)} = \mathsf{proj}_{\mathcal{Y}}(\mathbf{Y}_i^{(r)} - \alpha(\mathbf{A}_{\mathbf{Y}}^{(t)}\mathbf{Y}_i^{(r)} - \mathbf{Z}_{\mathbf{Y},i}^{(t)})) \tag{2.25}$$

where

$$\mathbf{Z}_{\mathbf{Y}}^{(t)} \triangleq (\mathbf{X}^{(t)})^T\mathbf{Z} + \rho(\mathbf{X}^{(t)})^T - (\mathbf{\Lambda}^{(t)})^T + \beta^{(t)}(\mathbf{Y}^{(t)})^T, \tag{2.26}$$

$$\mathbf{A}_{\mathbf{Y}}^{(t)} \triangleq (\mathbf{X}^{(t)})^T\mathbf{X}^{(t)} + (\rho + \beta^{(t)})\mathbf{I} \succ 0, \tag{2.27}$$

$\mathbf{Z}_{\mathbf{Y},i}$ denotes the ith column of matrix $\mathbf{Z}_{\mathbf{X}}$, α is the step size, which is chosen either as a constant $1/\lambda_{\max}(\mathbf{A}_{\mathbf{Y}}^{(t)})$, or by using some line search procedure (27); r denotes the iteration of the inner loop; for a given vector $\mathbf{w}$, $\mathsf{proj}_{\mathcal{Y}}(\mathbf{w})$ denotes the projection of it to the feasible set of $\mathbf{Y}_i$, which can be evaluated in closed-form (73, pp. 80) as follows

$$\mathbf{w}^+ = \mathsf{proj}_+(\mathbf{w}) \triangleq \max\{\mathbf{w}, \mathbf{0}_{K\times 1}\}, \tag{2.28}$$

$$\begin{aligned}\mathbf{Y}_i &= \mathsf{proj}_{\|\mathbf{w}^+\|_2^2 \le \tau}(\mathbf{w}^+) \\ &\triangleq \sqrt{\tau}\mathbf{w}^+ / \max\{\sqrt{\tau}, \|\mathbf{w}^+\|_2\}.\end{aligned} \tag{2.29}$$

Clearly, other algorithms such as the accelerated version of the gradient projection (74) can also be used to solve the $\mathbf{Y}$-subproblem. Here we pick GP for its simplicity.

In particular, it is worth noting that when $\mathbf{Z}$ is a sparse matrix, the complexity of computing $\mathbf{Z}\mathbf{Y}^{(t+1)}$ in (2.23) and $(\mathbf{X}^{(t)})^T\mathbf{Z}$ in (2.26) is only proportional to the number of nonzero entries of $\mathbf{A}$.

2.4 Numerical Results

In this section, we compare the proposed algorithm with a few existing SymNMF solvers on both synthetic and real data sets. We run each algorithm with 20 random initializations (except

for SNMF, which does not require external initialization). The entries of the initialized $\mathbf{X}$ (or $\mathbf{Y}$) follow *i.i.d.* uniform distribution in the range $[0, \tau]$. All algorithms are started with the same initial point each time, and all tests are performed using Matlab on a computer with Intel Core i5-5300U CPU running at 2.30GHz with 8GB RAM. Since the compared algorithms have different computational complexity, we use the objective values versus CPU time for fair comparison. We next describe different SymNMF solvers that are compared in our work.

2.4.1 Algorithms Comparison

In our numerical simulations, we compare the following algorithms.

Projected Gradient Descent (PGD) and Projected Newton method (PNewton) (54; 11) The PGD and PNewton directly use the gradient of the objective function. The key difference between them is that PGD adopts the identity matrix as a scaling matrix while PNewton exploits reduced Hessian for accelerating the convergence rate. The PGD algorithm converges slowly if the step size is not well selected, while the PNewton algorithm has high per-iteration complexity compared with the ANLS and NS-SymNMF, due to the requirement of computing the Hessian matrix at each iteration. Note that to the best of our knowledge, neither PGD nor PNewton possesses convergence or rate of convergence guarantees.

Alternating Non-negative Least Square (ANLS) (11) The ANLS method is a very competitive SymNMF solver, which can be implemented in parallel easily. ANLS reformulates SymNMF as

$$\min_{\mathbf{X},\mathbf{Y}\geq 0} g(\mathbf{X},\mathbf{Y}) = \|\mathbf{X}\mathbf{Y}^T - \mathbf{Z}\|_F^2 + \nu\|\mathbf{X}-\mathbf{Y}\|_F^2 \tag{2.30}$$

where $\nu > 0$ is the regularization parameter. One of shortcomings is that there is no theoretical guarantee that the ANLS method can converge to the set of KKT points of (1.1) or even producing two symmetric factors, although certain penalty terms for the difference between the factors ($\mathbf{X}$ and $\mathbf{Y}$) is included in the objective.

Symmetric Non-negative Matrix Factorization (SNMF) (10) The SNMF algorithm transforms the original problem to another one under the assumption that $\mathbf{Z}$ can be exactly decomposed by $\mathbf{X}\mathbf{X}^T$. Although SNMF often converges quickly in practice, there has been no theoretical analysis under the general case where $\mathbf{Z}$ cannot be exactly decomposed.

Coordinate Descent (CD) (55) The CD method updates each entry of $\mathbf{X}$ in a cyclic way. For updating each entry, we only need to find the roots of a fourth-order univariate function. However, CD may not converge to the set of KKT points of SymNMF. Instead, there is an additional condition given in (55) for checking whether the generated sequence converges to a unique limit point. A heuristic method for checking the condition is additionally provided, which requires, e.g., plotting the norm between the different iterates.

The Proposed NS-SymNMF The update rules of NS-SymNMF is similar to that of ANLS. The differences between them are that NS-SymNMF uses one additional block for dual variables and ANLS adds a penalty term. The dual update involved in NS-SymNMF benefits the convergence of the algorithm to KKT points of SymNMF.

We remark that in the implementation of NS-SymNMF we let $\tau = \max_k \theta_k$ (cf. (2.6)) and the maximum number of iterations of GP be 40. Also, we gradually increase the value of ρ from an initial value to meet condition (2.11) for accelerating the convergence rate (75). Here, the choice of ρ follows $\rho^{(t+1)} = \min\{\rho^{(t)}/(1-\epsilon/\rho^{(t)}), 6.1N\tau\}$ where $\epsilon = 10^{-3}$ as suggested in (76). We choose $\rho^{(1)} = \bar{\tau}$ for the case that $\mathbf{Z}$ can be exactly decomposed and $\sqrt{N}\bar{\tau}$ for the rest of cases, where $\bar{\tau}$ is the mean of $\theta_k, \forall k$. The similar strategy is also applied for updating $\beta^{(t)}$. We choose $\beta^{(t)} = 6\xi^{(t)}\|\mathbf{X}^{(t)}\mathbf{Y}^{(t)} - \mathbf{Z}\|_F^2/\rho^{(t)}$ where $\xi^{(t+1)} = \min\{\xi^{(t)}/(1-\epsilon/\xi^{(t)}), 1\}$ and $\xi^{(1)} = 0.01$, and only update $\beta^{(t)}$ once every 100 iterations to save CPU time. To update $\mathbf{Y}$, we implement the block pivoting method (49) since such method is faster than the GP method for solving the nonnegative least squares problem. If $\|\mathbf{Y}_i^{(t+1)}\|_2^2 \leq \tau$ is not satisfied, then we switch to GP on $\mathbf{Y}_i^{(t)}$. We also remark that we set the step size of PGD as 10^{-5} for all tested cases, and use the Matlab codes of PNewton and ANLS from http://math.ucla.edu/~dakuang/.

2.4.2 Performance on Synthetic Data

First we describe the two synthetic data sets that we have used in the first part of the numerical result.

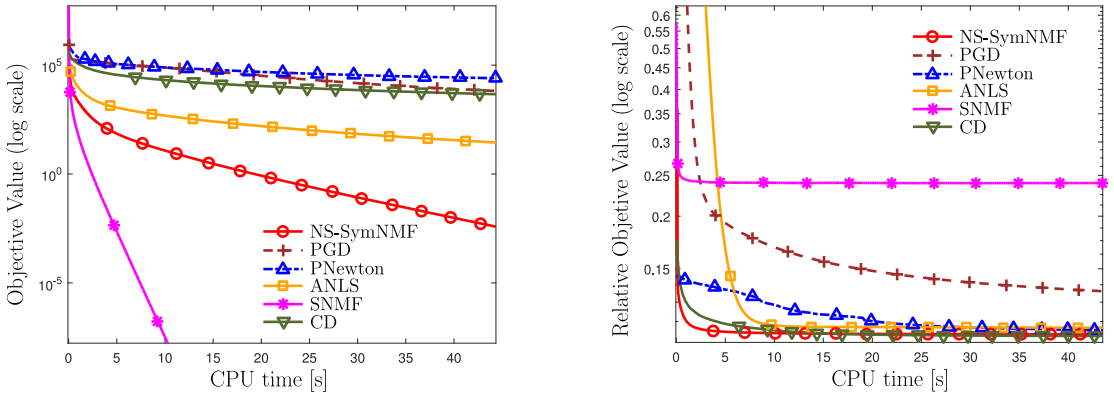
Data set I (Random symmetric matrices): We randomly generate two types of symmetric matrices, one is of low rank and the other is of full rank.

For the low rank matrix, we first generate a matrix $\mathbf{M}$ with dimension $N \times K$, whose entries follow *i.i.d.* Gaussian distribution. We use $\mathbf{M}_{i,j}$ to denote the (i,j)th entry of $\mathbf{M}$. Then generate a new matrix $\widetilde{\mathbf{M}}$ whose (i,j)th entry is $|\mathbf{M}_{i,j}|$. Finally, we obtain a positive symmetric $\mathbf{Z} = \widetilde{\mathbf{M}}\widetilde{\mathbf{M}}^T$ as the given matrix to be decomposed.

For the full rank matrix, we first randomly generate a $N \times N$ matrix, denoted as $\mathbf{P}$, whose entries follow *i.i.d.* uniform distribution in the interval $[0,1]$. Then we compute $\mathbf{Z} = (\mathbf{P} + \mathbf{P}^T)/2$.

Data set II (Adjacency matrices): One important application of SymNMF is graph partitioning, where the adjacency matrix of a graph is factorized. We randomly generate a graph as follows. First, fix the number of nodes to be N and the number of cluster to be 4, and the numbers of nodes within each cluster are $300, 500, 800, 400$. Second, we randomly generate data points whose relative distance will be used to construct the adjacency matrix. Specifically, data points $\{x_i\} \in \mathbb{R}$, $i = 1, \ldots, N$, are generated in one dimension. Within one cluster, data points follow *i.i.d.* Gaussian distribution. The means of the random variables in these 4 clusters are $2, 3, 6, 8$, respectively, and the variance is 0.5 for all distributions. Construct the similarity matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$, whose entries are determined by the Gaussian function $\mathbf{A}_{i,j} = \exp(-(x_i - x_j)^2/(2\sigma^2))$ where $\sigma^2 = 0.5$.

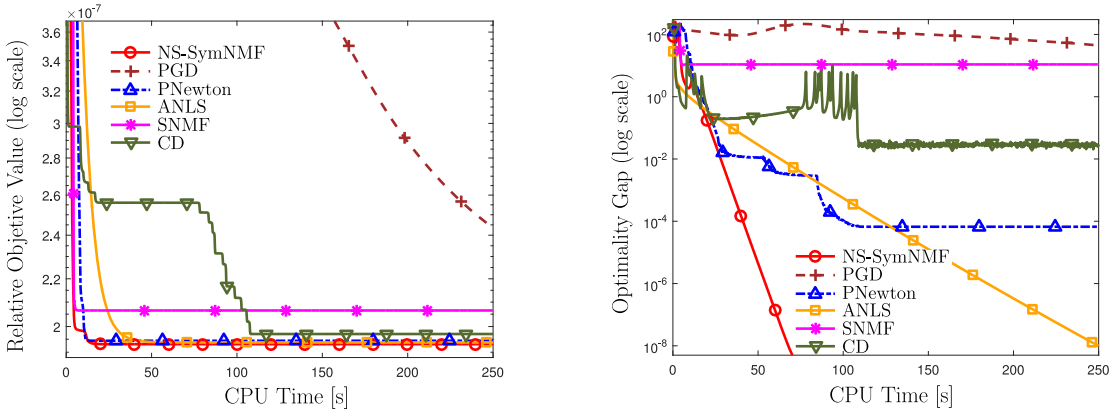
The convergence behaviors of different SymNMF solvers for the synthetic data sets are shown in Figure 2.1 and Figure 2.2. The results shown are averaged over 20 Monte Carlo (MC) trials with independently generated data. In Figure 2.1(a), the generated $\mathbf{Z}$ can be exactly decomposed by SymNMF. It can be observed that NS-SymNMF and SNMF converge to the global optimal solution quickly, and SNMF is the fastest one among all compared algorithms. However, the case where the matrix can be exactly factorized is not common in most practical applications. Hence, we also consider the case where the matrix $\mathbf{Z}$ cannot be factorized exactly by a $N \times K$ matrix.



(a) $N = 500$, $K = 60$. (b) $N = 500$, $K = 60$, and $\mathbf{Z}$ is full rank.

Figure 2.1 Data Set I: the convergence behaviors of different SymNMF solvers.

The results are shown in Figure 2.1(b) and we use the relative objective value for comparison, i.e., $\|\mathbf{X}\mathbf{X}^T - \mathbf{Z}\|_F^2 / \|\mathbf{Z}\|_F^2$. We can observe that NS-SymNMF and CD can achieve a lower objective value than other methods. It is worth noting that there is a gap between SNMF and others, since the assumption of SNMF is not satisfied in this case.



(a) Objective Value (b) Optimality Gap

Figure 2.2 Data Set II: the convergence behaviors of different SymNMF solvers; $N = 2000$, $K = 4$.

We also implement the algorithms on adjacency matrices (data set II), where the results are shown in Figure 2.2. The NS-SymNMF and SNMF algorithms converge very fast, but it can be observed that there is still a gap between SNMF and NS-SymNMF as shown in Figure 2.2(a).

We further show the convergence rates with respective to optimality gap versus CPU time in Figure 2.2(b). The optimality gap (2.14) measures the closeness between the generated sequence and the true stationary point. To get rid of the effect of the dimension of $\mathbf{Z}$, we use $\|\mathbf{X} - \mathsf{proj}_+[\mathbf{X} - \nabla_{\mathbf{X}}(g(\mathbf{X}, \mathbf{Y}))]\|_\infty$ as the optimality gap. It is interesting to see the "swamp" effect (77), where the objective value generated by the CD algorithm remains almost constant during the time period from around 25s to 75s although actually the corresponding iterates do not converge, and then the objective value starts decreasing again.

Table 2.1 Local Optimality

N	$\lambda_{\min}(\mathcal{T})$	δ	Local Optimality (true)
50	2.71×10^{-4}	0.42	100%
100	4.16×10^{-4}	0.37	100%
500	1.8×10^{-2}	0.91	100%

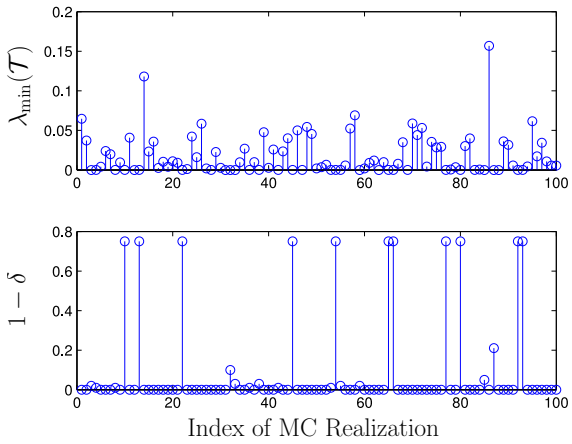


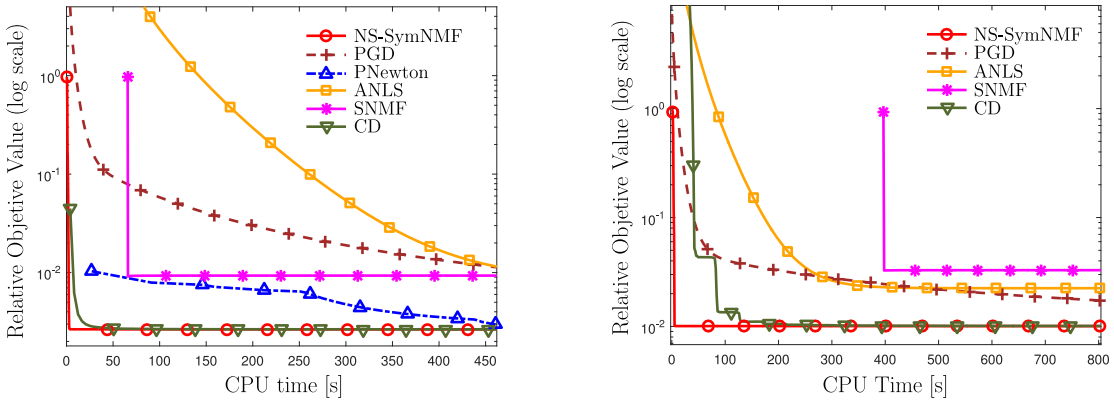
Figure 2.3 Checking local optimality condition, where $N = 500$.

2.4.3 Checking Global/Local Optimality

After the NS-SymNMF algorithm has converged, the local/global optimality can be checked according to Theorem 3 and Theorem 4. To find an appropriate δ that satisfying the condition where $\lambda_{\min}(T) > 0$, we initialize δ as 1 and decrease it by 0.01 each time and check the minimum eigenvalue of $\mathcal{T}$. Here, we use data set II with the fixed ratio of the number of nodes within each

cluster (i.e., $3:5:8:4$) and test on the different total numbers of nodes. The simulation results are shown in Table 2.1 with 100 Monte Carlo trials, where the average value of $\lambda_{\min}(\mathcal{T})$ and δ are given. Further, the percentage of being able to find a valid $\delta > 0$ that ensures $\lambda_{\min}(T) > 0$ is listed as the last column. It can be observed that there always exists a δ such that $\mathcal{T}$ is positive definite in all cases that we have tested. This indicates that (with high probability) the proposed algorithm converges to a locally optimal solution. In Figure 2.3, we provide the values of δ that make the corresponding $\lambda_{\min}(\mathcal{T}) > 0$ at each realization.

We also remark that in practice we stop the algorithm in finite steps, so only an approximate KKT point will be obtained, and the degree of such approximation can be measured by the optimality gap defined in (2.14).



(a) Mean of the objective values: Reuters data set

(b) Mean of the objective values: TDT2 data set

Figure 2.4 The convergence behaviors of different SymNMF solvers for the dense similarity matrix.

2.4.4 Performance on Real Data

We also implement the algorithm on a few real data sets in clustering applications, which will be described in the next paragraphs.

Dense Similarity Matrix:

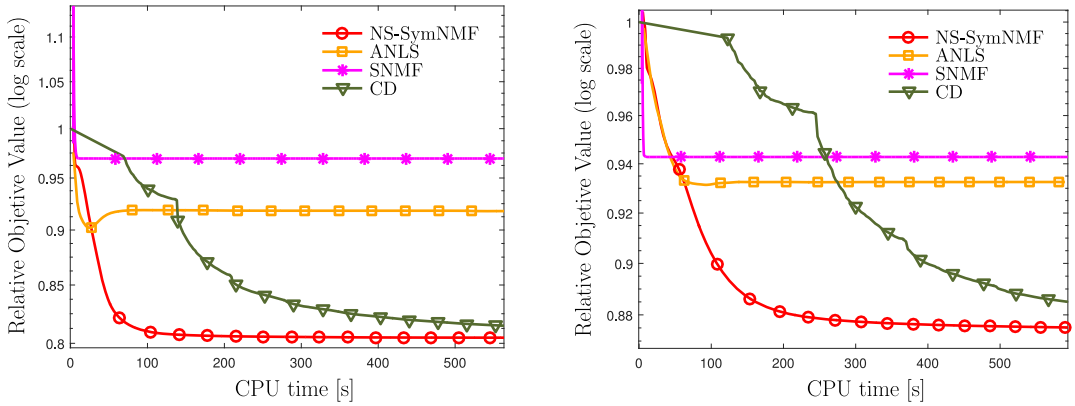
Table 2.2 Mean and Standard Deviation of $\|\mathbf{X}\mathbf{X}^T - \mathbf{Z}\|_F^2/\|\mathbf{Z}\|_F^2$ Obtained by the Final Solution of Each Algorithm based on Random Initializations (dense similarity matrices)

Dense Data Sets	Reuters (78)	TDT2 (78)
N	4,633	8,939
K	25	25
NS-SymNMF	**2.65e-3±3.31e-10**	**1.01e-2±5.35e-9**
PGD (54)	1.14e-2±1.18e-5	1.74e-2±7.34e-6
PNewton (54)	2.98e-3±3.71e-6	-
ANLS (11)	1.16e-2±1.61e-5	2.25e-2±1.25e-6
SNMF (10)	9.32e-3	3.29e-2
CD (55)	2.66e-3±2.04e-8	1.01e-2±1.21e-6

We generate the dense similarity matrices based on the two real data sets: Reuters-21578 (78) and TDT2 (78). We use the 10th subset of the processed Reuters-21578 data set, which includes $N = 4,633$ documents divided into $K = 25$ classes. The number of features is 18,933. Topic detection and tracking 2 (TDT2) corpus includes two newswires (APW and NYT), two radio programs (VOA and PRI) and two television programs (CNN and ABC). We use the 10th subset of the processed TDT2 data set with $K = 25$ classes which includes $N = 8,939$ documents and each of them has 36,771 features. We comment that the 10th TDT2 subset is the largest among the all TDT2 and Reuters subsets. Any other subset can be used equally well. The similarity matrix is constructed by the Gaussian function where the difference between two documents is measured by all features using the Euclidean distance (78).

The means and standard deviations of the objective values of the final solutions are shown in Table 2.2. Convergence results of the algorithms are shown in Figure 2.4. For the Reuters and TDT2 datasets, before SNMF completes the eigenvalue decomposition for the first iteration, CD and NS-SymNMF have already obtained very low objective values. Also, since the calculation of Hessian in PNewton is time consuming for large scale matrices, the result of PNewton is out of range in Figure 2.4(b).

Sparse Similarity Matrix: We also generate multiple convergence curves for each algorithm with random initializations based on some sparse real data sets.



(a) Mean of the objective values: email-Enron data set

(b) Mean of the objective values: loc-Brightkite data set

Figure 2.5 The convergence behaviors of different SymNMF solvers for the sparse similarity matrix.

Table 2.3 Mean and Standard Deviation of $\|\mathbf{X}\mathbf{X}^T - \mathbf{Z}\|_F^2/\|\mathbf{Z}\|_F^2$ Obtained by the Final Solution of Each Algorithm based on Random Initializations (sparse similarity matrices)

Sparse Data Sets	email-Enron (79)	loc-Brightkite (80)
N	36,692	58,228
K	50	50
#nonzero	367,662	428,156
NS-SymNMF	**8.05e-1±4.66e-4**	**8.75e-1±9.52e-4**
ANLS (11)	9.18e-1±6.20e-3	9.33e-1±1.93e-3
SNMF (10)	9.69e-1	9.43e-1
CD (55)	8.13e-1±1.47e-3	8.84e-1±1.49e-3

Email-Enron network data set (79): Enron email corpus includes around half million emails. We use the relationships between two email addresses to construct the similarity matrix for decomposing. If an address i sent at least one email to address j, then we take $\mathbf{A}_{i,j} = \mathbf{A}_{j,i} = 1$. Otherwise, we set $\mathbf{A}_{i,j} = \mathbf{A}_{j,i} = 0$.

Brightkite data set (80): Brightkite was a location-based social networking website. Users were able to share their current locations by checking-in. The friendships of the users were maintained by Brightkite. The way of constructing the similarity matrix is the same as the Enron email data set.

The means and standard deviations of the objective values of the final solutions are shown in Table 2.3. From the simulation results shown in Figure 2.5, it can be observed that the NS-SymNMF algorithm converges faster than CD, while SNMF and ANLS converge to some points where the relative objective values are higher than the one obtained by NS-SymNMF.

CHAPTER 3. STOCHASTIC SYMMETRIC NONNEGATIVE MATRIX FACTORIZATION

3.1 Introduction

In practical problems when there are multiple samples obtained, stochastic-type algorithm is the one of the most efficient options of handling stochastic optimization problems. Recently, the stochastic projected gradient descent (SPGD) methods are proposed for dealing with stochastic nonconvex problems (81; 82). However, there has been no convergence guarantee when directly applying SPGD to solve the stochastic SymNMF problem, since there is no global Lipschitz continuity of the gradient of the objection function. Classical stochastic approximation methods can also be used, but without convergence and rate of convergence guarantees. Fast convergence rates of stochastic ADMM algorithms are presented recently (83; 84), however, these algorithms only work for stochastic convex optimization problems. In fact, none of the works has rigorous theoretical justification that they can be applied directly for SymNMF in the stochastic settings.

The most relevant algorithm that uses the nonconvex splitting method for solving SymNMF was proposed in (85), but the algorithm, called NS-SymNMF, only works for the case where the given data is deterministic. In this chapter, we consider the stochastic setting of matrix factorization that potentially make the SymNMF more practical. The proposed algorithm, named stochastic nonconvex splitting SymNMF (SNS-SymNMF), is a generalization of the previous NS-SymNMF algorithm, which is able to factorize the realizations of the random observation matrix in each iteration. Further, actually the convergence proof of NS-SymNMF does not apply to that of SNS-SymNMF, since the iterates are coupled with the random data matrices as the algorithm proceeds such that the boundness of the iterates is not clear if the convergence proof of NS-SymNMF was used.

In this work, SNS-SymNMF is proposed for problem (1.1), where the underlying distribution is unknown, but realizations of $\mathbf{Z}$ are available sequentially. The proposed algorithm belongs to the class of stochastic algorithms, because at each iteration only a few samples of the observation matrix are used. Based on different ways in which the samples are utilized, we analyze the performance of the algorithm in terms of its convergence rates to the set of stationary solutions of problem (1.1). The main contributions of this chapter are given below.

- The proposed algorithm possesses sublinear convergence rate guarantees. When an aggregate of the past samples is used (possibility with non-uniform weighting), the algorithm converges sublinearly to the stationary points of problem (1.1) in mean-square; when the instantaneous samples are used, the algorithm converges sublinearly to a neighborhood around the stationary solutions. To our best knowledge, this is the first stochastic algorithm that can possess a sublinear convergence rate for stochastic SymNMF.
- We demonstrate the performance of the proposed stochastic algorithm for clustering problems. It is shown that SNS-SymNMF is much faster compared with some existing algorithms for generic stochastic nonconvex optimization problems numerically. Further, due to the use of non-uniform aggregate sampling, the proposed algorithm is capable of tracking changes of the community structure.

3.2 Stochastic Nonconvex Splitting for SymNMF

3.2.1 Main Assumptions

The sequentially sampled data $\widehat{\mathbf{Z}}^{(i)}$ are assumed to be independent and identically distributed (*i.i.d.*) realizations of the random matrix $\mathbf{Z}$, where i denotes the index of the sample. Rather than assuming the unbiased gradient and bounded variance of the stochastic gradient in most stochastic gradient methods (82), we only need to make assumptions on samples for SymNMF. Specifically, we assume the following.

- A1) Unbiased sample: $\mathbb{E}[\widehat{\mathbf{Z}}^{(i)}] = \overline{\mathbf{Z}} \quad \forall i$;

Table 3.1 Rules of Aggregating Samples

Mini-batch	Aggregate	Weighted Aggregate
$\mathbf{Z}_1^{(t)} = \frac{1}{L}\sum_{i=(t-1)L+1}^{tL} \widehat{\mathbf{Z}}_1^{(i)}$	$\mathbf{Z}_1^{(t)} = \frac{1}{t}\sum_{i=1}^{t} \widehat{\mathbf{Z}}_1^{(i)}$	$\mathbf{Z}_1^{(t)} = \frac{2}{t(t+1)}\sum_{i=1}^{t} i\widehat{\mathbf{Z}}_1^{(i)}$
$\mathbf{Z}_2^{(t)} = \frac{1}{L}\sum_{i=(t-1)L+1}^{tL} \widehat{\mathbf{Z}}_2^{(i)}$	$\mathbf{Z}_2^{(t)} = \frac{1}{t}\sum_{i=1}^{t} \widehat{\mathbf{Z}}_2^{(i)}$	$\mathbf{Z}_2^{(t)} = \frac{2}{t(t+1)}\sum_{i=1}^{t} i\widehat{\mathbf{Z}}_2^{(i)}$

- A2) Bounded variance: $\mathsf{Tr}[\mathsf{Var}[\widehat{\mathbf{Z}}^{(i)}]] = \mathbb{E}[\|\widehat{\mathbf{Z}}^{(i)} - \overline{\mathbf{Z}}\|_F^2] \leq \sigma^2 \quad \forall i;$
- A3) Bounded magnitude: $\|\widehat{\mathbf{Z}}^{(i)}\|_F \leq \mathcal{Z} < \infty \quad \forall i.$

In practice, the magnitude of samples is finite, so A3 is valid (11; 82).

3.2.2 The Problem Formulation for Stochastic SymNMF

We start by considering the following reformulation of problem (1.1) to the following problem:

$$\begin{aligned} \min_{\mathbf{X},\mathbf{Y}} \quad & \frac{1}{2}\|\mathbf{X}\mathbf{Y}^T - \mathbb{E}_{\mathbf{Z}}[\mathbf{Z}]\|_F^2 \\ \text{s.t.} \quad & \mathbf{X} = \mathbf{Y},\ \mathbf{Y} \geq 0, \|\mathbf{Y}_i\|_2^2 \leq \tau,\ \forall\ i \end{aligned} \tag{3.1}$$

where $\mathbf{Z}$ is a symmetric matrix; $\tau > 0$ is some given constant.

Under A1, it is easy to check that when τ is sufficiently large (with a lower bound dependent on $\overline{\mathbf{Z}}$), then problem (3.1) is *equivalent* to problem (1.1), in the sense that there is a one-to-one correspondence between the stationary points of problem (1.1) and (3.1), where the stationary condition of problem (1.1) is given by (27, Proposition 2.1.2)

$$\left\langle \left(\mathbf{X}^*(\mathbf{X}^*)^T - \overline{\mathbf{Z}}\right)\mathbf{X}^*, \mathbf{X} - \mathbf{X}^* \right\rangle \geq 0,\ \forall \mathbf{X} \in \mathcal{X}.$$

where $\mathbf{X}^*$ denotes the stationary points. To be precise, we have the following result.

Lemma 3. *Let $\overline{\mathbf{Z}}_{i,k}$ denote the (i,k)th entry of the matrix $\overline{\mathbf{Z}}$. Under A1 – A3, suppose $\tau > \theta_k, \forall k$ where*

$$\theta_k \triangleq \frac{\overline{\mathbf{Z}}_{k,k} + \sqrt{\sum_{i=1}^{N} \overline{\mathbf{Z}}_{i,k}^2}}{2}, \tag{3.2}$$

then a point $\mathbf{X}^*$ is a stationary point of problem (1.1) if and only if $\mathbf{X}^*$ is a stationary point of problem (3.1).

Although the objective function does not have Lipschitz continuous gradient, Theorem 3 suggests that we can solve (1.1) within a compact set.

3.2.3 The Framework of SNS for SymNMF

To this end, let us construct the augmented Lagrangian for (2.3), given by

$$\mathcal{L}(\mathbf{X},\mathbf{Y};\mathbf{\Lambda}) = \frac{1}{2}\|\mathbf{X}\mathbf{Y}^T - \bar{\mathbf{Z}}]\|_F^2 + \langle \mathbf{Y}-\mathbf{X}, \mathbf{\Lambda}\rangle + \frac{\rho}{2}\|\mathbf{Y}-\mathbf{X}\|_F^2 \tag{3.3}$$

where $\mathbf{\Lambda} \in \mathbb{R}^{N\times K}$ is a matrix of dual variables (or Lagrange multipliers); $\langle\cdot\rangle$ denotes the inner product operator; $\rho > 0$ is a penalty parameter whose value will be determined later.

The proposed SNS-SymNMF algorithm alternates between the primal updates of variables $\mathbf{X}$ and $\mathbf{Y}$, and the dual update for $\mathbf{\Lambda}$. We split the data samples into two groups where $\widehat{\mathbf{Z}}_1^{(i)}$ is used for updating $\mathbf{Y}$ and $\widehat{\mathbf{Z}}_2^{(i)}$ is used for $\mathbf{X}$, respectively. Our algorithm is also capable of dealing with a few different ways of aggregating the samples at each iteration:

1. A Mini-Batch of L instantaneous samples are used;
2. An aggregate of the historical samples is used;
3. A special weighted aggregate of the historical samples is used.

See Table 3.1 for their mathematical descriptions. In the table, t denotes the tth iteration of the algorithm; $\mathbf{Z}_1^{(t)}$ and $\mathbf{Z}_2^{(t)}$ are the actual (aggregated) samples used in our algorithm.

In the following, we provide the main steps of the proposed algorithm. The implementation of each step will be provided shortly. At iteration $t+1$, we first compute the objective value evaluated at the previous sample, followed by the primal updates for $\mathbf{X}$ and $\mathbf{Y}$, finally the dual variable $\mathbf{\Lambda}$ is updated. Specifically,

$$\beta^{(t)} = \frac{8}{\rho}\|\mathbf{X}^{(t)}(\mathbf{Y}^{(t)})^T - \mathbf{Z}_2^{(t-1)}\|_F^2, \tag{3.4a}$$

$$\mathbf{Y}^{(t+1)} = \arg\min_{\mathbf{Y}\geq 0, \|\mathbf{Y}_i\|_2^2\leq\tau,\forall i} \widehat{\mathcal{L}}_{\mathbf{Y}}(\mathbf{X}^{(t)},\mathbf{Y};\mathbf{\Lambda}^{(t)};\mathbf{Z}_1^{(t)}), \tag{3.4b}$$

$$\mathbf{X}^{(t+1)} = \arg\min_{\mathbf{X}} \widehat{\mathcal{L}}_{\mathbf{X}}(\mathbf{X},\mathbf{Y}^{(t+1)};\mathbf{\Lambda}^{(t)};\mathbf{Z}_2^{(t)}), \tag{3.4c}$$

$$\mathbf{\Lambda}^{(t+1)} = \mathbf{\Lambda}^{(t)} + \rho(\mathbf{X}^{(t+1)} - \mathbf{Y}^{(t+1)}) \tag{3.4d}$$

where we have defined

$$\widehat{\mathcal{L}}_{\mathbf{Y}}(\mathbf{X}^{(t)},\mathbf{Y};\mathbf{\Lambda}^{(t)};\mathbf{Z}_1^{(t)}) \triangleq \frac{1}{2}\|\mathbf{X}^{(t)}\mathbf{Y}^T-\mathbf{Z}_1^{(t)}\|_F^2+\frac{\rho}{2}\|\mathbf{X}^{(t)}-\mathbf{Y}+\mathbf{\Lambda}^{(t)}/\rho\|_F^2+\frac{\beta^{(t)}}{2}\|\mathbf{Y}-\mathbf{Y}^{(t)}\|_F^2,$$
$$\widehat{\mathcal{L}}_{\mathbf{X}}(\mathbf{X},\mathbf{Y}^{(t+1)};\mathbf{\Lambda}^{(t)};\mathbf{Z}_2^{(t)}) \triangleq \frac{1}{2}\|\mathbf{X}(\mathbf{Y}^{(t+1)})^T-\mathbf{Z}_2^{(t)}\|_F^2+\frac{\rho}{2}\|\mathbf{X}-\mathbf{Y}^{(t+1)}+\mathbf{\Lambda}^{(t)}/\rho\|_F^2.$$

We remark that using independent samples for the $\mathbf{X}$ and $\mathbf{Y}$ update is critical in the convergence analysis of the algorithm.

3.2.4 Implementation of the SNS-SymNMF Algorithm

The implementation of SNS-SymNMF is shown in Algorithm 1. The updates of variable $\mathbf{X}$ and $\mathbf{Y}$ in each subproblem are the similar as the way in NS-SymNMF but with different strategy of using samples.

The SNS-SymNMF Algorithm. Leveraging the efficient calculation of $\mathbf{Y}^{(t+1)}$ and $\mathbf{X}^{(t+1)}$ (see (2.24) and (2.25)), we summarize the algorithm as shown in Algorithm 1, where T denotes the total number of iterations.

Algorithm 1 The SNS-SymNMF Algorithm

1: **Input:** $\mathbf{Y}^{(1)}$, $\mathbf{X}^{(1)}$, $\mathbf{\Lambda}^{(1)}$, and ρ
2: **for** $t=1,\ldots,T$ **do**
3: Update $\beta^{(t)}$ according to (3.4a)
4: Select data using Table 3.1
5: Update $\mathbf{Y}^{(t+1)}$ by solving (3.4b)
6: Update $\mathbf{X}^{(t+1)}$ by solving (3.4c)
7: Update $\mathbf{\Lambda}^{(t+1)}$ using (3.4d)
8: **end for**
9: **Output**: Iterate $\mathbf{Y}^{(r)}$ chosen uniformly random from $\{\mathbf{Y}^{(t)}\}_{t=1}^T$.

3.3 Convergence Analysis

The convergence analysis is built upon a series of lemmas (shown in the supplemental materials of (86)), which characterize the relationship among the augmented Lagrangian, the primal/dual variables as well as the random samples.

We also remark the convergence proof of SNS-SymNMF is different from the work in the previous chapter. Here we start from the proof of the boundness of the $\mathbf{X}$-iterate, then the convergence of the algorithm to stationary points can be characterized.

Theoretical Results. First, when a mini-batch of samples are used at each iteration, we have the following result.

Theorem 5. *Suppose A1 – A3 hold true. Then the iterates generated by the SNS-SymNMF algorithm with Mini-Batch samples satisfy the following relation*

$$\mathbb{E}[\mathcal{P}_{\textit{Mini-Batch}}(\mathbf{X}^{(r)}, \mathbf{Y}^{(r)}, \mathbf{\Lambda}^{(r)})] \leq \frac{1}{T}\mathcal{C}(\mathcal{U} + \frac{\sigma^2}{L}) + \frac{\mathcal{W}\sigma^2}{L}$$

where $\mathcal{C}$,$\mathcal{U}$,$\mathcal{W}$ are some constants.

Theorem 5 says that using the Mini-Batch samples the SNS-SymNMF algorithm converges sublinearly to a ball of size $\mathcal{W}\sigma^2/L$ around the stationary points of problem (2.3). Further, the radius of the ball can be reduced when increasing the number of samples L.

Second, if all the past samples are averaged using the same weight, then the algorithm can converge to the stationary points of the stochastic SymNMF problem.

Theorem 6. *Suppose A1 – A3 hold true and the following is satisfied*

$$\rho > 8NK\tau^2. \tag{3.5}$$

Then the following statements are true for SNS-SymNMF with averaged samples:

1. *The equality constraint is satisfied in the limit, i.e.,*

$$\lim_{t\to\infty} \mathbb{E}[\|\mathbf{X}^{(t)} - \mathbf{Y}^{(t)}\|_F^2] \to 0.$$

2. *The sequence $\{\mathbf{X}^{(t)}, \mathbf{Y}^{(t)}, \mathbf{\Lambda}^{(t)}\}$ is bounded, and every limit point of the sequence is a stationary point of problem (3.1).*

Below we show that the gap $\mathbb{E}[\mathcal{P}(\mathbf{X}^{(r)}, \mathbf{Y}^{(r)}, \mathbf{\Lambda}^{(r)})]$ goes to zero in mean-square sublinearly.

Theorem 7. *Suppose A1 – A3 hold true. Then the iterates generated by the SNS-SymNMF algorithm with aggregate samples satisfy the following relation*

$$\mathbb{E}[\mathcal{P}_{\mathit{aggregate}}(\mathbf{X}^{(r)}, \mathbf{Y}^{(r)}, \mathbf{\Lambda}^{(r)})] \leq \frac{\mathcal{C}\mathcal{S} + \mathcal{C}\sigma^2 + \mathcal{K}\sigma^2}{T}$$

where $\mathcal{C}, \mathcal{S}, \mathcal{K}$ are some constants.

Theorem 6 and Theorem 7 show that the stochastic SymNMF can converge to a stationary point of (3.1) in mean-square, and in a sublinear manner. Then, we have the following corollary directly.

Corollary 2. *Suppose A1 – A3 hold true. Then the iterates generated by the SNS-SymNMF algorithm with weighted aggregate samples satisfy the following relation*

$$\mathbb{E}[\mathcal{P}_{\mathit{weighted}}(\mathbf{X}^{(r)}, \mathbf{Y}^{(r)}, \mathbf{\Lambda}^{(r)})] \leq \frac{\mathcal{C}\mathcal{S} + \mathcal{C}\sigma^2 + \mathcal{K}'\sigma^2}{T}$$

where $\mathcal{K}' \geq \mathcal{K}$.

Remark 1. Those constants, such as $\mathcal{C}, \mathcal{U}, \mathcal{W}, \mathcal{S}, \mathcal{K}$, mentioned in the theorems are only dependent on the initialization of the algorithm and parameters of given problems, such as $N, K, \tau, \mathcal{Z}$. The explicit expressions of the constants can be found in the supplemental materials of (86).

It is worth noting that when $\sigma^2 = 0$, our convergence analysis of the SNS-SymNMF algorithm still holds true for the deterministic case (85).

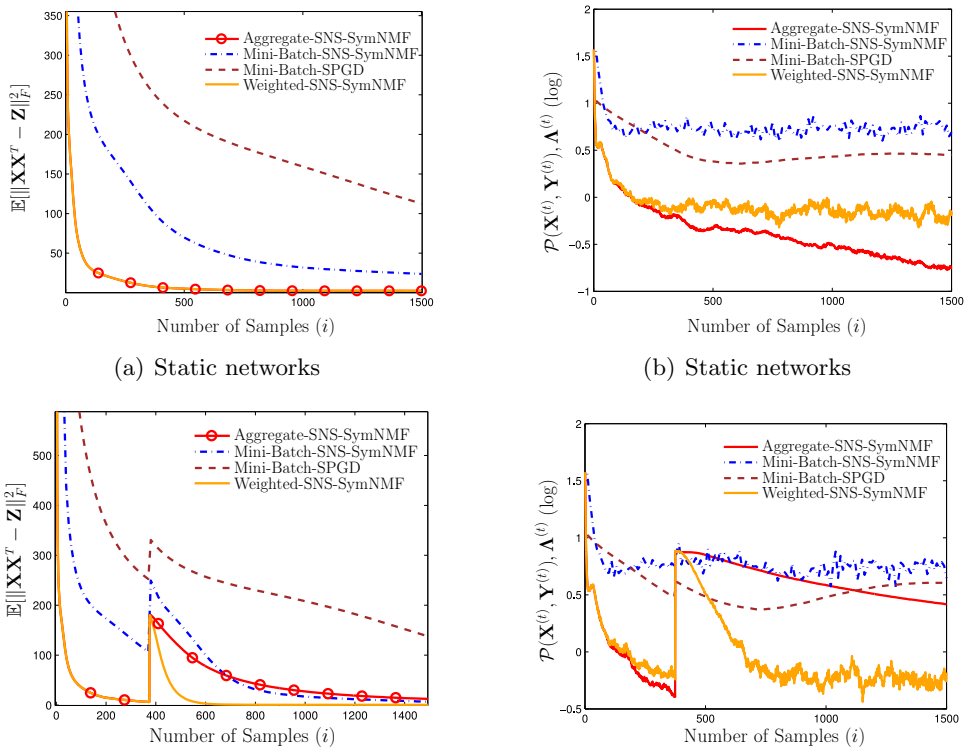
We also remark that given a required error, when the dimension of the problems increases, the stochastic algorithms need a more total number of iterations to achieve this error.

3.4 Numerical Results

3.4.1 Synthetic Data Set

Data Set Description. We use a similar random graph as adopted in (14) for spectral clustering. The graph is generated as follows. For each time slot, data points $\{x_i\} \in \mathbb{R}$, $i = 1, \ldots, N$, are generated in one dimension. We specify 4 clusters. The numbers of data points in each cluster

are 12, 24, 48 and 36. Within each cluster, data points follow an *i.i.d.* Gaussian distribution. The means of the random variables in these 4 clusters are $2, 4, 6, 8$, respectively, and the variance is 0.5 for all distributions. Then, construct the similarity matrix $\widehat{\mathbf{Z}}_1^{(i)} \in \mathbb{R}^{N\times N}$ (or $\widehat{\mathbf{Z}}_2^{(i)}$), whose (i,j)th entry is determined by the Gaussian function $\exp(-(x_i - x_j)^2/(2\sigma^2))$ where $\sigma^2 = 0.5$. Finally, we repeat the process mentioned above to generate a series of adjacency matrices for the community detection problem. The mean of the adjacency matrix represents the ground truth of the connections among the nodes and variance measures the uncertainty of each sample. Based on this model, we know that the weights between two points which belong to the same cluster are very likely higher than the weights between two points which belong to different clusters.



(a) Static networks (b) Static networks

(c) Dynamic networks (d) Dynamic networks

Figure 3.1 The convergence behaviors. The parameters are $K = 4$; $N = 120$; $L = 10$. The x-axis represents the total number of observed samples.

Algorithms Comparison. Each point in Figure 3.1 is an average of 20 independent Monte Carlo (MC) trials. All algorithms are started with the same initial point each time, and the entries of

the initialized $\mathbf{X}$ (or $\mathbf{Y}$) follow an *i.i.d.* uniform distribution in the range $[0, \tau]$. Mini-Batch SPGD (82) is applied to solve problem (3.1) where the step-size α is 0.01. Note that this algorithm cannot be directly applied to solve problem (1.1) due to the lack of Lipschitz continuous gradient. The proposed SNS-SymNMF uses two groups of data at each iteration, while Mini-Batch-SPGD only needs one. For fair comparison, in the simulation Mini-Batch-SPGD uses $(\mathbf{Z}_1^{(t)} + \mathbf{Z}_2^{(t)})/2$ as the input sample. Also, when the Mini-Batch strategy is used, the algorithms perform updates every L independent samples, where L is fixed.

We remark that in the implementation of SNS-SymNMF we let $\tau = \max_k \theta_k$, and gradually increase the value of ρ from an initial value to meet condition (3.5) for accelerating the convergence rate (75). Here, the choice of ρ follows $\rho^{(t+1)} = \min\{\rho^{(t)}/(1 - \epsilon/\rho^{(t)}), 8.1NK\tau^2\}$ where $\epsilon = 10^{-3}$ as suggested in (76), and $\rho^{(1)} = N\tau$. To update $\mathbf{Y}$, we use the block pivoting method (49).

The SNS-SymNMF algorithm is performed using different data sampling rules. From Figure 3.1(a), it is shown that the aggregate-SNS-SymNMF algorithm converges faster than Mini-Batch-SPGD and Mini-Batch-SNS-SymNMF since the variance of samples is reduced by the aggregated data. The weighted-SNS-SymNMF algorithm is slightly slower than aggregate-SNS-SymNMF, but still presents a sublinear convergence rate. As shown in Figure 3.1(b), the optimality gap plateaus in Mini-Batch-SNS-SymNMF and Mini-Batch-SPGD due to the sample aggregation rules, which is consistent with the theoretical analysis shown in Theorem 5. The optimality gap of Mini-Batch-SNS-SymNMF is larger than that of Mini-Batch-SPGD, since the number of samples used for each block is only a half of Mini-Batch-SPGD. Here, to get rid of the effect of the dimension of $\mathbf{Z}$, we use $\|\mathbf{X} - \mathsf{proj}_+[\mathbf{X} - \nabla_{\mathbf{X}}(f(\mathbf{X}))]\|_\infty$ as the optimality gap, where proj_+ denotes the nonnegative projection operator.

The convergence behaviors for dynamic networks are shown in Figure 3.1(c) and Figure 3.1(d), where the means of the random variables in the 4 clusters are changed to $1, 7, 3, 5$ at the 400th sample. Aggregate-SNS-SymNMF performs worse than weighted-SNS-SymNMF because of the aggregated errors. Although Mini-Batch-SNS-SymNMF and Mini-Batch-SPGD can adapt to the network topology variation, constant optimality gaps still remain as can be observed in Figure 3.1(d).

For the weighted-SNS-SymNMF algorithm, since more weights are given to the current data samples, the change of the network topology can be tracked. Therefore, weighted-SNS-SymNMF can still give a very low objective value after the 400th sample compared with other algorithms.

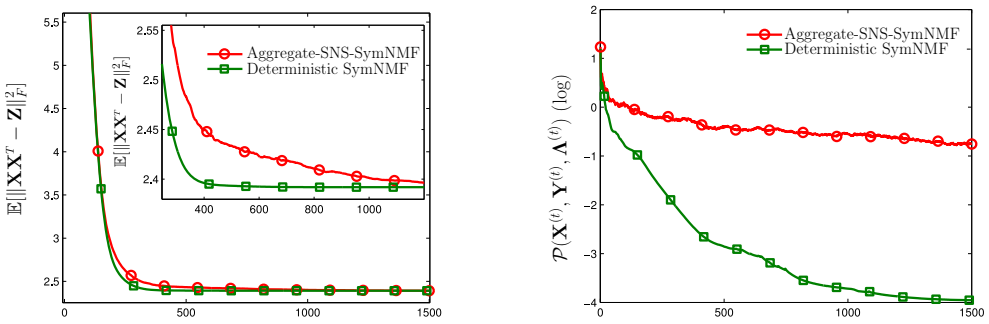


Figure 3.2 The convergence behaviors. The parameters are $K = 4$; $N = 120$; $L = 10$. The x-axis represents the total number of the observed samples for stochastic SymNMF and iterations for deterministic SymNMF.

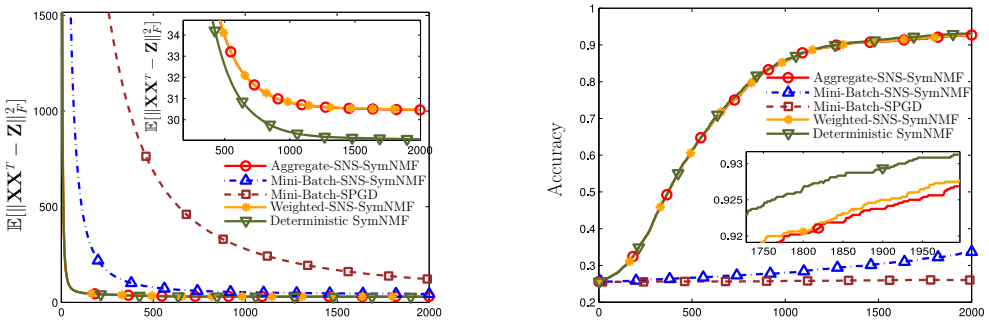


Figure 3.3 The convergence behaviors. The parameters are $K = 5$; $N = 240$; $L = 10$. The x-axis represents the total number of observed samples for stochastic SymNMF and iterations for deterministic SymNMF.

We also compare the performance of the SNS-SymNMF algorithm and the deterministic SymNMF algorithm where the samples are replaced by $\bar{\mathbf{Z}}$ in SNS-SymNMF. The results are shown in Figure 3.2. It can be observed that the SNS-SymNMF algorithm has a similar convergence rate with NS-SymNMF in terms of the objective values. However, deterministic SymNMF has a faster convergence rate than SNS-SymNMF with respective to the optimality gap, which is expected, since deterministic SymNMF uses the mean of the adjacency matrix without any uncertainty.

3.4.2 Real Data Set

Data Set Description. we use the 6th subset of the processed topic detection and tracking (TDT2) data set with 10 classes[1] which includes 3050 documents and each of them has 36771 features. The adjacency matrix is constructed by the self-tuning method (87), where the weight between the ith sample and the jth one is given by $w_{i,j} = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|_2^2/(\sigma_i\sigma_j)), \forall i \neq j$. The local scale σ_i is computed by the Euclidean distance between $\mathbf{x}_i$ and its $\widehat{k}$th neighbor, where $\mathbf{x}_i$ denotes the ith document vector which is normalized to have unit 2-norm and $i = 1, \ldots, N$. We use $\widehat{k} = 7$ as suggested in (87) and enforce $w_{i,i} = 0, \forall i$. Then the (i, j)th entry of the similarity matrix $\widehat{\mathbf{Z}}_1^{(i)}$ (or $\widehat{\mathbf{Z}}_2^{(i)}$) is computed as in the normalized cut (14) which is $d_i^{-1/2} w_{i,j} d_j^{-1/2}$ where $d_i = \sum_{i'}^{N} w_{i,i'}, \forall i'$.

In order to mimic the stochastic setting, we select 5 classes that have larger number of documents than the others in the 6th subset of TDT2. The total numbers of documents in these 5 classes are 1843, 440, 226, 144, and 103. Then, for each time slot, we uniformly pick up 100, 50, 45, 15, 30 documents from the selected 5 classes to form $\widehat{\mathbf{Z}}_1^{(i)}$, and then independently perform the same sampling process again to form $\widehat{\mathbf{Z}}_2^{(i)}$. The average of all samples is considered as the true mean (i.e., $\overline{\mathbf{Z}}$) for NS-SymNMF. The variance of samples in this case is $\sigma^2 = 32.32$.

Algorithms Comparison. The simulation results shown in Figure 3.3 are based on 20 MC trials. It can be observed that Mini-Batch algorithms converge slowly compared with aggregated/weighted SNS-SymNMF and NS-SymNMF, since Mini-Batch algorithms only use a subset of samples. Although NS-SymNMF shows a lower objective value than SNS-SymNMF, it is interesting to see that SNS-SymNMF has a similar convergence rate as NS-SymNMF in terms of the objective values with only a small difference. Furthermore, the accuracy obtained by NS-SymNMF and aggregated/weighted SNS-SymNMF is only slightly different during the whole process as the algorithms proceed. Therefore, the new variant of SymNMF, SNS-SymNMF, can be considered as an online

[1] see http://www.cad.zju.edu.cn/home/dengcai/Data /TextData.html.

algorithm that deals with clustering problems, which is not only processing the real-time data sequentially but also can provide accurate clustering results[2].

Finally, we remark that the previous literatures (11; 10) have already shown the advantages of deterministic SymNMF in terms of clustering accuracy compared with classic methods, such as K-means variants, NMF variants, spectral clustering variants. Here, we focus on the stochastic setting for SymNMF and omit the accuracy results for other methods.

We also remark that in this work we just adopt a very simple version of Mini-Batch methods. The main purpose is to take the Mini-Batch methods as the counterparts for the average/weighted aggregation rules and to show the impact of the variance of samples on performance of algorithms. Actually, there is a tradeoff on selecting the length L as the Mini-Batch algorithm proceeds. A more reasonable way of choosing L is discussed in (82) and more variants of Mini-Batch algorithms for stochastic SymNMF could be considered as the future work.

[2]More simulations related to the computational time, impact of sample variance, and parameter tuning are shown in the supplemental materials of (86), where the numerical results with larger networks are also included.

CHAPTER 4. PERTURBED ALTERNATING GRADIENT DESCENT

4.1 Introduction

Many recent works have been focused on the performance analysis and/or design of algorithms with convergence guarantees to local minimum points/SS2 for nonconvex optimization problems. These include the trust region method (88), cubic regularized Newton's method (89; 90), and a mixed approach of the first-order and seconde-order methods (91), etc. However, these algorithms typically require second-order information, therefore they incur high computational complexity when problem dimension becomes large.

There has been a line of work on stochastic gradient descent algorithms, where properly scaled Gaussian noise is added to the iterates of the gradient at each time [also known as stochastic gradient Langevin dynamics, (SGLD)]. Some theoretical works have pointed out that SGLD not only converges to the local minimum points asymptotically but also may escape from local minima (92; 93). Unfortunately, these algorithms require a large number of iterations with $\mathcal{O}(1/\epsilon^4)$ steps to achieve the optimal point. There are fruitful results that show some carefully designed algorithms can escape from strict saddle points efficiently, such as negative-curvature-originated-from noise (Neon) (94), Neon2 (95), Neon$^+$(96) and gradient descent with one-step escaping (GOSE) (97). The Neon-type of algorithms utilizes the stochastic first-order updates to find the negative curvature direction, and GOSE just needs one negative curvature descent step with calculation of eigenvectors when the iterates of the algorithm are near the saddle point for saving the computational burden.

On the other hand, there is also a line of work analyzing the deterministic GD type method. With random initializations, it has been shown that GD only converges to SS2 for unconstrained smooth problems (98). More recently, block coordinate descent, block mirror descent and proximal block coordinate descent have been proven to almost always converge to SS2 with random initializations (99), but there is no convergence rate reported. Unfortunately, a follow-up study indicated

that GD requires exponential time to escape from saddle points for certain pathological problems (100). Adding some noise occasionally to the iterates of the algorithm is another way of finding the negative curvature. A perturbed version of GD has been proposed with convergence guarantees to SS2 (101), which shows a faster provable convergence rate than the ordinary gradient descent algorithm with random initializations. Furthermore, the accelerated version of PGD (PAGD) is also proposed in (102), which shows the fastest convergence rate among all Hessian free algorithms.

Table 4.1 Convergence rates of algorithms to SS2 with the first order information, where $p \geq 4$, and $\widetilde{\mathcal{O}}$ hides factor ploylog(d).

Algorithm	Iterations	(ϵ, γ)-SS2
SGD (44)	$\mathcal{O}(d^p/\epsilon^4)$	$(\epsilon, \epsilon^{1/4})$
SGLD (92)	$\mathcal{O}(d^p/\epsilon^4)$	$(\epsilon, \epsilon^{1/2})$
Neon+SGD (94)	$\widetilde{\mathcal{O}}(1/\epsilon^4)$	$(\epsilon, \epsilon^{1/2})$
Neon+Natasha (94)	$\widetilde{\mathcal{O}}(1/\epsilon^{13/4})$	$(\epsilon, \epsilon^{1/4})$
Neon2+SGD (95)	$\widetilde{\mathcal{O}}(1/\epsilon^4)$	$(\epsilon, \epsilon^{1/2})$
Neon$^+$ (96)	$\widetilde{\mathcal{O}}(1/\epsilon^{7/4})$	$(\epsilon, \epsilon^{1/2})$
PGD (101)	$\widetilde{\mathcal{O}}(1/\epsilon^2)$	$(\epsilon, \epsilon^{1/2})$
PAGD (102)	$\widetilde{\mathcal{O}}(1/\epsilon^{7/4})$	$(\epsilon, \epsilon^{1/2})$
PA-GD/PA-PP (This work)	$\widetilde{\mathcal{O}}(1/\epsilon^{7/3})$	$(\epsilon, \epsilon^{1/3})$

4.1.1 Scope of This Work

In this chapter, we consider a smooth unconstrained optimization problem, and develop a perturbed AGD algorithm (PA-GD) which converges (with high probability) to the set of SS2 with a global sublinear rate (103). Our work is inspired by the works (101; 44), which developed novel perturbed GDs that escapes from strict saddle points. Similarly as in (101), we also divide the entire iterates of GD into three types of points: those whose gradients are large, those that are local minimum, and those that are strict saddle points. At a given point, when the size of the gradient is large enough, we just implement the ordinary AGD. When the gradient norm is small, which may be either strict saddle or local minimum, a perturbation will be added on the iterates to help to escape from the saddle points.

From the above section, we know that many works have been developed to make use of negative curvature information around the saddle points. Unfortunately, these techniques cannot be directly applied to the BCD/AGD- type of algorithms. The *key challenge* here is that at each iteration only part of the variables are updated, therefore we have access only to partial second order information at the points of interest. For example, consider a quadratic objective function shown in Figure 4.1. While fixing one block, the problem is strongly convex with respect to the other block, but the entire problem is nonconvex. Even if the iterates converge for each block to the minimum points within the block, the stationary point could still be a saddle point for the overall objective function. Therefore, the analysis of how AGD type of algorithms exploit the negative curvature is one of the main tasks in this chapter.

To the best of our knowledge, there is no work on modifying AGD algorithms to escape from strict saddle points with any convergence rate. The main contributions of this work are as follows.

4.1.2 Contributions

In this work, we design and analyze a perturbed AGD algorithm for solving an unconstrained nonconvex problem, namely perturbed AGD. Through the perturbation of AGD, the algorithm is guaranteed to converge to a set of SS2 of a nonconvex problem with high probability. By utilizing the matrix perturbation theory, convergence rate of the proposed algorithm is also established, which shows that the algorithm takes $\mathcal{O}(\text{polylog}(d)/\epsilon^{7/3})$ iterations to achieve an $(\epsilon, \epsilon^{1/3})$-SS2 with high probability. Also, considering the fact that there is a strong relation between GD and proximal point algorithm, we also study a perturbed alternating proximal point (PA-PP) algorithm with some random perturbation. By leveraging the techniques proposed in this work, we show that PA-PP, which may not need to calculate the gradient at each step, converges as fast as PA-GD in the order of ϵ . The comparison of the algorithms which only use the first order information for escaping from strict saddle points is summarized as shown in Table 4.1.

The main contributions of the work are highlighted below:

1. To the best of our knowledge, it is the first time that the convergence analysis shows that some variants of AGD (using first-order information) can converge to SS2 for nonconvex optimization problems.
2. The convergence rate of the perturbed AGD algorithm is analyzed, where the choice of the step size is only dependent on certain maximum Lipschitz constant over blocks rather than all variables. This is one of the major difference between GD and AGD.
3. By further extending the analysis in this work, we also show that PA-PP can also escape from the strict points efficiently with the speed of $\mathcal{O}(\text{polylog}(d)/\epsilon^{7/3})$.

4.2 Preliminaries

In this chapter, we use bold upper case letters without subscripts (e.g., $\mathbf{X}, \mathbf{Y}$) to denote matrices and bold lower case letters without subscripts (e.g., $\mathbf{x}, \mathbf{y}$) represent vectors. Notation $\mathbf{x}_k$ denotes the kth block of vector $\mathbf{x} \in \mathbb{R}^{d\times 1}$. We use $\nabla_k f(\mathbf{x}_{-k}, \mathbf{x}_k)$ to denote the partial gradient with respect to its kth block variable while the remaining one is fixed. Notation $\mathbb{B}_{\mathbf{x}}(r)$ denotes a d-dimensional ball centered at $\mathbf{x}$ with radius r, and $\lambda_{\min}(\mathbf{X})$, $\lambda_{\max}(\mathbf{X})$ denote the smallest and largest eigenvalues of matrix $\mathbf{X}$ respectively.

4.2.1 Definitions

The objective function has the following properties.

Definition 1. *A differentiable function $f(\cdot)$ is L-smooth with gradient Lipschitz constant L (uniformly Lipschitz continuous), if*

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|, \quad \forall \mathbf{x}, \mathbf{y}.$$

The function is called block-wise smooth with gradient Lipschitz constants $\{L_k\}$, if

$$\|\nabla_k f(\mathbf{x}_{-k}, \mathbf{x}_k) - \nabla_k f(\mathbf{x}_{-k}, \mathbf{x}'_k)\| \leq L_k\|\mathbf{x}_k - \mathbf{x}'_k\|, \ \forall \mathbf{x}, \mathbf{x}'$$

or with gradient Lipschitz constants $\{\widetilde{L}_k\}$*, if*

$$\|\nabla_k f(\mathbf{x}_{-k}, \mathbf{x}_k) - \nabla_k f(\mathbf{x}'_{-k}, \mathbf{x}_k)\| \leq \widetilde{L}_k \|\mathbf{x}_{-k} - \mathbf{x}'_{-k}\|, \ \forall \mathbf{x}, \mathbf{x}'.$$

Further, let $L_{\max} := \max\{L_k, \widetilde{L}_k, \forall k\} \leq L$.

Definition 2. *For a differentiable function* $f(\cdot)$*, if* $\|\nabla f(\mathbf{x})\| = 0$*, then* $\mathbf{x}$ *is a first-order stationary point. If* $\|\nabla f(\mathbf{x})\| \leq \epsilon$*, then* $\mathbf{x}$ *is an* ϵ*-first-order stationary point.*

Definition 3. *For a differentiable function* $f(\cdot)$*, if* $\mathbf{x}$ *is a SS1, and there exists* $\epsilon > 0$ *so that for any* $\mathbf{y}$ *in the* ϵ*-neighborhood of* $\mathbf{x}$*, we have* $f(\mathbf{x}) \leq f(\mathbf{y})$*, then* $\mathbf{x}$ *is a local minimum. A saddle point* $\mathbf{x}$ *is a SS1 that is not a local minimum. If* $\lambda_{\min}(\nabla^2 f(\mathbf{x})) < 0$*,* $\mathbf{x}$ *is a strict (non-degenerate) saddle point.*

Definition 4. *A twice-differentiable function* $f(\cdot)$ *is* ρ*-Hessian Lipschitz if*

$$\|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y})\| \leq \rho \|\mathbf{x} - \mathbf{y}\|, \quad \forall \mathbf{x}, \mathbf{y}. \tag{4.1}$$

Definition 5. *For a* ρ*-Hessian Lipschitz function* $f(\cdot)$*,* $\mathbf{x}$ *is a second-order stationary point if* $\|\nabla f(\mathbf{x})\| = 0$ *and* $\lambda_{\min}(\nabla^2 f(\mathbf{x})) \geq 0$*. If the following holds*

$$\|\nabla f(\mathbf{x})\| \leq \epsilon, \quad \textit{and} \quad \lambda_{\min}(\nabla^2 f(\mathbf{x})) \geq -\gamma \tag{4.2}$$

where $\epsilon, \gamma > 0$*, then* $\mathbf{x}$ *is a* (ϵ, γ)*-SS2.*

Assumption 1. *Function* $f(\cdot)$ *is L-smooth, block-wise smooth with gradient Lipschitz constants* $\{L_k, \widetilde{L}_k\}, k = 1, 2$*, and* ρ*-Hessian Lipschitz.*

4.3 Perturbed Alternating Gradient Descent

4.3.1 Algorithm Description

AGD is a classical algorithm that optimizes the variables of an optimization problem in an alternating manner (27), meaning that when one block of variables is updated, the remaining block

Algorithm 2 Perturbed Alternating Gradient Descent (PA-GD) $(\mathbf{x}^{(0)}, L_{\max}, L, \rho, \epsilon, \delta, \Delta f)$

Input: $\mathcal{P}_1 = (1 + \frac{L}{L_{\max}})$, $\mathcal{P}_2 = (1 + \frac{L\log(2d)}{L_{\max}})$, $\chi = 6\max\{\log(\frac{\mathcal{P}_1^6\mathcal{P}_2^2 d L_{\max}^{5/3}\Delta_f}{c^5\rho^{1/3}\epsilon^{7/3}\delta}, 4\}$, $\eta = \frac{c}{L_{\max}}$, $r = \frac{c^3}{\chi^3}\frac{\rho\epsilon}{L_{\max}\mathcal{P}_1^3\mathcal{P}_2}$, $g_{\mathsf{th}} = \frac{c^2\epsilon}{(\chi\mathcal{P}_1)^3\mathcal{P}_2}$, $f_{\mathsf{th}} = \frac{c^5\epsilon^2}{L_{\max}(\chi\mathcal{P}_1)^6\mathcal{P}_2^2}$, $t_{\mathsf{th}} = \frac{L_{\max}\chi\mathcal{P}_1}{c^2(L_{\max}\rho\epsilon)^{\frac{1}{3}}}$
for $t = 0, 1, \ldots$ **do**
  $\mathbf{x}_1^{(t+1)} = \mathbf{x}_1^{(t)} - \eta\nabla_1 f(\mathbf{x}_1^{(t)}, \mathbf{x}_2^{(t)})$
  if $\sum_{k=1}^2 \|\nabla_k f(\mathbf{h}_{-k}^{(t)}, \mathbf{x}_k^{(t)})\|^2 \le g_{\mathsf{th}}^2$ and $t - t_{\mathsf{p}} > t_{\mathsf{th}}$ **then**
    $\widetilde{\mathbf{x}}^{(t)} \leftarrow \mathbf{x}^{(t)}$ and $t_{\mathrm{p}} \leftarrow t$
    $\mathbf{x}^{(t)} = \widetilde{\mathbf{x}}^{(t)} + \xi^{(t)}$, $\xi^{(t)}$ uniformly taken from $\mathbb{B}_0(r)$
    $\mathbf{x}_1^{(t+1)} = \mathbf{x}_1^{(t)} - \eta\nabla_1 f(\mathbf{x}_1^{(t)}, \mathbf{x}_2^{(t)})$
  end if
  $\mathbf{x}_2^{(t+1)} = \mathbf{x}_2^{(t)} - \eta\nabla_2 f(\mathbf{x}_1^{(t+1)}, \mathbf{x}_2^{(t)})$
  if $t - t_{\mathsf{p}} = t_{\mathsf{th}}$ and $f(\mathbf{x}^{(t)}) - f(\widetilde{\mathbf{x}}^{(t_{\mathsf{p}})}) > -f_{\mathsf{th}}$ **then**
    return $\widetilde{\mathbf{x}}^{t_{\mathsf{p}}}$
  end if
end for

is fixed to be the same as its previous solution. Mathematically, the iterates of AGD are updated by the following rule

$$\mathbf{x}_k^{(t+1)} = \mathbf{x}_k^{(t)} - \eta\nabla_k f(\mathbf{h}_{-k}^{(t)}, \mathbf{x}_k^{(t)}), \quad k = 1, 2 \tag{4.3}$$

where superscript (t) denotes the iteration counter; $\mathbf{h}_{-1}^{(t)} := \mathbf{x}_2^{(t)}$ and $\mathbf{h}_{-2}^{(t)} := \mathbf{x}_1^{(t+1)}$; $\eta > 0$ is the step size. AGD can be considered as a special case of block coordinate gradient descent (29; 30).

Our proposed algorithm is based on AGD, but modified in a way similar to the recent work (101), which adds some noise in PGD. The details of the implementation of PA-GD are shown in Algorithm 2, where c is a constant so that $\eta = c/L_{\max}$, Δ_f denotes the difference of the objective value at the initial point and global optimal solution, ϵ represents the predefined target error.

In each update of variables, we implement one step of the block gradient descent, and then proceed to the next block. Once the algorithm has sufficient decrease of the objective value, it implies that the algorithm converges to some good solution. Otherwise, some perturbation may be needed to help the iterates escape from the saddle points. If after the perturbation the objective value does not decrease sufficiently after a number of further iterations, the algorithm terminates and returns the iterate before the last perturbation.

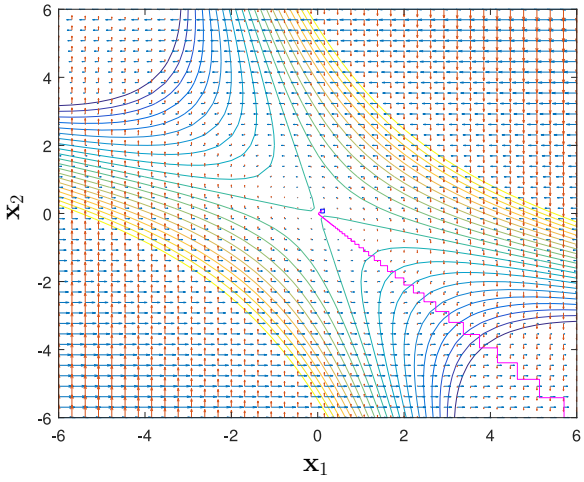


Figure 4.1 Contour of the objective values and the trajectory (pink color) of PA-GD started near strict saddle point $[0, 0]$. The objective function is $f(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x}, \mathbf{x} = [\mathbf{x}_1; \mathbf{x}_2] \in \mathbb{R}^{2\times 1}$ where $\mathbf{A} := [1 \quad 2; 2 \quad 1] \in \mathbb{R}^{2\times 2}$, and the length of the arrows indicate the strength of $-\nabla f(\mathbf{x})$ projected onto directions $\mathbf{x}_1, \mathbf{x}_2$.

To illustrate the practical behavior of the algorithm, we provide an example that shows the trajectory of AGD after a small perturbation at a stationary point. In Figure 4.1, it is clear that $\mathbf{x} = [0; 0]$ is a SS1 and also a strict saddle point since the eigenvalues of $\mathbf{A}$ are -1 and 3 respectively. When $\mathbf{x}_1$ is fixed, function $f(\mathbf{x})$ is convex with respect to $\mathbf{x}_2$ and vice versa, however, the objective function is nonconvex. It can be observed that PA-GD can escape from the strict saddle point efficiently.

4.3.2 Convergence Rate Analysis

Despite the fact that PA-GD exploits a different way of updating variables, we will show that it can still escape from strict saddle points with high probability with suitable perturbation. The main theorem is presented as follows.

Theorem 8. *Under Assumption 1, there exists a constant* $c_{\max}$ *such that: for any* $\delta \in (0, 1]$, $\epsilon \leq \frac{L_{\max}^2}{\rho}$, $\Delta_f := f(\mathbf{h}_{-1}^{(0)}, \mathbf{x}_1^{(0)}) - f^*$, *and constant* $c \leq c_{\max}$, *with probability* $1 - \delta$, *the iterates*

generated by PA-GD converge to an ϵ-SS2 $\mathbf{x}$ satisfying

$$\|\nabla f(\mathbf{x})\| \leq \epsilon, \quad \textit{and} \quad \lambda_{\min}(\nabla^2 f(\mathbf{x})) \geq -(L_{\max}\rho\epsilon)^{1/3}$$

in the following number of iterations:

$$\mathcal{O}\left(\frac{L_{\max}^{5/3}\mathcal{P}_1^7\mathcal{P}_2^2\Delta f}{\rho^{1/3}\epsilon^{7/3}}\log^7\left(\frac{\mathcal{P}_1^6\mathcal{P}_2^2 d L_{\max}^{5/3}\Delta_f}{c^5\rho^{1/3}\epsilon^{7/3}\delta}\right)\right) \tag{4.4}$$

where f^* denotes the global minimum value of the objective function, and $\mathcal{P}_1 = (1 + L/L_{\max})$ and $\mathcal{P}_2 = (1 + L\log(2d)/L_{\max})$.

Remark 2. When $\eta = c_{\max}/L$ is used, the convergence rate of PA-GD is

$$\mathcal{O}\left(\frac{L_{\max}^{5/3}\log^2(2d)\Delta f}{\rho^{1/3}\epsilon^{7/3}}\log^7\left(\frac{\mathcal{P}_1^6\mathcal{P}_2^2 d L_{\max}^{5/3}\Delta_f}{c^5\rho^{1/3}\epsilon^{7/3}\delta}\right)\right). \tag{4.5}$$

It shows that if a smaller step size is used, the convergence rate of PA-GD is faster (with smaller constants) since the linear dependency of $\mathcal{P}_1^7$ and $\mathcal{P}_2^2$ in (4.4) both disappear. This property is consistent with the known result when BCD is used in convex optimization problems, i.e., when a smaller step size is used, the rate could become better; e.g., see (104, Theorem 2.1).

4.4 Perturbed Alternating Proximal Point

In many applications, AGD may not be efficient in the sense that the convergence rate of gradient in each block may be very slow. For example, consider matrix factorization problem $\min_{\mathbf{X},\mathbf{Y}} \|\mathbf{Z} - \mathbf{X}\mathbf{Y}\|_F^2$ where $\mathbf{Z} \in \mathbb{R}^{m\times d}$ is the given data, $d \gg m$, and $\mathbf{X} \in \mathbb{R}^{m\times r}, \mathbf{Y} \in \mathbb{R}^{r\times d}$ are two block variables. For this problem, the alternating least squares algorithm (which exactly minimizes each block) would be a faster algorithm compared with the AGD which only uses gradient steps.

In this section, we consider the classical proximal point algorithm (105) in which each block of variables is exactly minimized with respect to certain quadratic surrogate. To be specific, we can replace (4.3) in Algorithm 2 by

$$\mathbf{x}_k^{(t+1)} = \arg\min_{\mathbf{x}_k} f(\mathbf{h}_{-k}^{(t)}, \mathbf{x}_k) + \frac{\nu}{2}\|\mathbf{x}_k - \mathbf{x}_k^{(t)}\|^2, \ k = 1, 2 \tag{4.6}$$

Algorithm 3 Perturbed Alternating Proximal Point (PA-PP) $(\mathbf{x}^{(0)}, L_{\max}, L, \rho, \epsilon, \delta, \Delta f)$

Input: $\mathcal{P} = (1 + \frac{L\log(2d)}{L_{\max}})$, $\chi = 6\max\{\log(\frac{\mathcal{P}^2 d L_{\max}^{5/3}\Delta_f}{c^5\rho^{1/3}\epsilon^{7/3}\delta}, 4\}$, $\nu = \frac{L_{\max}}{c}$, $r = \frac{c^3}{\chi^3}\frac{\rho\epsilon}{L_{\max}\mathcal{P}}$, $g_{\mathsf{th}} = \frac{c^2\epsilon}{\chi^3\mathcal{P}}$, $f_{\mathsf{th}} = \frac{c^5\epsilon^2}{L_{\max}\chi^6\mathcal{P}^2}$, $t_{\mathsf{th}} = \frac{L_{\max}\chi}{c^2(L_{\max}\rho\epsilon)^{\frac{1}{3}}}$
for $t = 0, 1, \ldots$ **do**
  for $k = 1, 2$ **do**
    $\mathbf{x}_k^{(t+1)} = \arg\min_{\mathbf{x}_k} f(\mathbf{h}_{-k}^{(t)}, \mathbf{x}_k) + \frac{\nu}{2}\|\mathbf{x}_k - \mathbf{x}_k^{(t)}\|^2$
  end for
  if $\|\mathbf{x}^{(t+1)} - \mathbf{x}^{(t)}\| \le g_{\mathsf{th}}/\nu$ and $t - t_{\mathsf{p}} > t_{\mathsf{th}}$ **then**
    $\widetilde{\mathbf{x}}^{(t)} \leftarrow \mathbf{x}^{(t)}$ and $t_{\mathrm{p}} \leftarrow t$
    $\mathbf{x}^{(t)} = \widetilde{\mathbf{x}}^{(t)} + \xi^{(t)}$, $\xi^{(t)}$ uniformly taken from $\mathbb{B}_0(r)$
    for $k = 1, 2$ **do**
      $\mathbf{x}_k^{(t+1)} = \arg\min_{\mathbf{x}_k} f(\mathbf{h}_{-k}^{(t)}, \mathbf{x}_k) + \frac{\nu}{2}\|\mathbf{x}_k - \mathbf{x}_k^{(t)}\|^2$
    end for
  end if
  if $t - t_{\mathsf{p}} = t_{\mathsf{th}}$ and $f(\mathbf{x}^{(t)}) - f(\widetilde{\mathbf{x}}^{(t_{\mathsf{p}})}) > -f_{\mathsf{th}}$ **then**
    return $\widetilde{\mathbf{x}}^{t_{\mathsf{p}}}$
  end if
end for

where $\nu > 0$ is penalty parameter. The iteration can be explicitly written as

$$\mathbf{x}_k^{(t+1)} = \mathbf{x}_k^{(t)} - \frac{1}{\nu}\nabla_k f(\mathbf{h}_{-k}^{(t)}, \mathbf{x}_k^{(t+1)}), \qquad k = 1, 2, \tag{4.7}$$

which has the similar form as the PA-GD algorithm, but with the step size being $\eta := 1/\nu$, and with gradient evaluated at the new iterate. The resulting algorithm, detailed in the table above, is referred to as the perturbed alternating proximal point (PA-PP). It is worth noting that when the subproblem is convex, such as $\min_{\mathbf{X},\mathbf{Y}} \|\mathbf{Z} - \mathbf{X}\mathbf{Y}\|_F^2$, ν only needs to be a small number to make the corresponding subproblem strongly convex. This property is useful in practice.

Next, we can also give the convergence rate of PA-PP.

Corollary 3. *Under Assumption 1, there exists a constant $c_{\max}$ such that: for any $\delta \in (0, 1]$, $\epsilon \le \frac{L_{\max}^2}{\rho}$, $\Delta_f := f(\mathbf{h}_{-1}^{(0)}, \mathbf{x}_1^{(0)}) - f^*$, and constant $c \le c_{\max}$, with probability $1 - \delta$, the iterates generated by PA-PP converges to an ϵ-SS2 $\mathbf{x}$ satisfying*

$$\|\nabla f(\mathbf{x})\| \le \epsilon, \quad \textit{and} \quad \lambda_{\min}(\nabla^2 f(\mathbf{x})) \ge -(L_{\max}\rho\epsilon)^{1/3}$$

in the following number of iterations:

$$\mathcal{O}\left(\frac{L_{\max}^{5/3}\mathcal{P}^2\Delta f}{\rho^{1/3}\epsilon^{7/3}}\log^7\left(\frac{\mathcal{P}^2 d L_{\max}^{5/3}\Delta_f}{c^5\rho^{1/3}\epsilon^{7/3}\delta}\right)\right)$$

where f^* denotes the global minimum value of the objective function, and $\mathcal{P} = (1+L\log(2d)/L_{\max})$.

Comparing with Theorem 8, we can find that term $\mathcal{P}_1^7, \mathcal{P}_1 > 2$ is removed so the convergence rate of PA-PP is slightly faster than PA-GD.

4.5 Convergence Analysis

In this section, we will present the main proof steps of convergence analysis of PA-GD.

4.5.1 The Main Difficulty of the Proof

Gradient Descent: GD searches the descent direction of the objective function in the entire space $\mathbb{R}^d$. Without loss of generality, we assume $\mathbf{x}^{(0)} = 0$. According to the mean value theorem, the GD update can be expressed as

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \eta\nabla f(\mathbf{x}^{(t)}) = \mathbf{x}^{(t)} - \eta\nabla f(0) - \eta\left(\int_0^1 \nabla^2 f(\theta\mathbf{x}^{(t)})d\theta\right)\mathbf{x}^{(t)}. \tag{4.8}$$

It can be observed that the update rule of GD contains the information of the Hessian matrix at point $\mathbf{x}^{(t)}$, i.e., $\nabla^2 f(\theta\mathbf{x}^{(t)})$. To be more specific, letting $\mathbf{H} \triangleq \nabla^2 f(\mathbf{x}^*)$ where $\mathbf{x}^*$ denotes an ϵ-SS2 satisfying (4.2), we can rewrite (4.8) as

$$\mathbf{x}^{(t+1)} = (\mathbf{I} - \eta\mathbf{H})\mathbf{x}^{(t)} - \eta\Delta^{(t)}\mathbf{x}^{(t)} - \eta\nabla f(0) \tag{4.9}$$

where $\Delta^{(t)} := \int_0^1(\nabla^2 f(\theta\mathbf{x}^{(t)}) - \mathbf{H})d\theta$.

Based on the ρ-Hessian Lipschitz property, we can quantify $\|\Delta^{(t)}\|$ that is upper bounded by the difference of iterates. By exploiting the negative curvature of the Hessian matrix at saddle point $\mathbf{x}^*$, we can project the iterate onto the direction $\vec{d}$ where the eigenvalue of $\mathbf{I} - \eta\mathbf{H}$ is greater than 1. This leads to the fact that the norm of the iterates projected along direction $\vec{d}$ will be increasing exponentially as the algorithm proceeds around point $\mathbf{x}^*$, implying the sequence generated by GD is escaping from the saddle point. The details of characterizing the convergence rate have been analyzed previously in (101).

Alternating Gradient Descent: However, the AGD algorithm only updates partial variables of vector $\mathbf{x}$, which belong to a subspace of the feasible set. Similarly, from the mean value theorem we can express the AGD rule of updating variables with assuming $\mathbf{x}^{(0)} = 0$ as follows:

$$\begin{aligned}\mathbf{x}^{(t+1)} &= \mathbf{x}^{(t)} - \eta \begin{bmatrix} \nabla_1 f(\mathbf{x}_1^{(t)}, \mathbf{x}_2^{(t)}) \\ \nabla_2 f(\mathbf{x}_1^{(t+1)}, \mathbf{x}_2^{(t)}) \end{bmatrix} \\ &= \mathbf{x}^{(t)} - \eta \nabla f(0) - \eta \int_0^1 \mathbf{H}_l^{(t)} d\theta \mathbf{x}^{(t+1)} - \eta \int_0^1 \mathbf{H}_u^{(t)} d\theta \mathbf{x}^{(t)} \end{aligned} \tag{4.10}$$

where

$$\mathbf{H}_l^{(t)} := \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \nabla_{21}^2 f(\theta \mathbf{x}_1^{(t+1)}, \theta \mathbf{x}_2^{(t)}) & \mathbf{0} \end{bmatrix} \quad \text{and} \quad \mathbf{H}_u^{(t)} := \begin{bmatrix} \nabla_{11}^2 f(\theta \mathbf{x}_1^{(t)}, \theta \mathbf{x}_2^{(t)}) & \nabla_{12}^2 f(\theta \mathbf{x}_1^{(t)}, \theta \mathbf{x}_2^{(t)}) \\ \mathbf{0} & \nabla_{22}^2 f(\theta \mathbf{x}_1^{(t+1)}, \theta \mathbf{x}_2^{(t)}) \end{bmatrix}.$$

From the above expression, it can be seen clearly that the update rule of AGD does not include a full Hessian matrix at any point but only partial ones. Furthermore, the right hand side of (4.10) not only contains the second order information of the previous point, i.e., $[\mathbf{x}_1^{(t)}, \mathbf{x}_2^{(t)}]$ but also the one of the most recently updated point, i.e., $[\mathbf{x}_1^{(t+1)}, \mathbf{x}_2^{(t)}]$. These represent the main challenges in understanding the behavior of the sequence generated by the AGD algorithm.

4.5.2 The Main Idea of the Proof

Although the second order information is divided into two parts, we can still characterize the recursion of the iterates around strict saddle points. We can also split $\mathbf{H}$ as two parts, which are

$$\mathbf{H}_u = \begin{bmatrix} \nabla_{11}^2 f(\mathbf{x}^*) & \nabla_{12}^2 f(\mathbf{x}^*) \\ \mathbf{0} & \nabla_{22}^2 f(\mathbf{x}^*) \end{bmatrix}, \quad \mathbf{H}_l = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \nabla_{21}^2 f(\mathbf{x}^*) & \mathbf{0} \end{bmatrix}, \tag{4.11}$$

and obviously we have $\mathbf{H} = \mathbf{H}_l + \mathbf{H}_u$.

Then, recursion (4.10) can be written as

$$\mathbf{x}^{(t+1)} + \eta \mathbf{H}_l \mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \eta \mathbf{H}_u \mathbf{x}^{(t)} - \eta \Delta_u^{(t)} \mathbf{x}^{(t)} - \eta \Delta_l^{(t)} \mathbf{x}^{(t+1)} \tag{4.12}$$

where $\Delta_u^{(t)} := \int_0^1 (\mathbf{H}_u^{(t)}(\theta) - \mathbf{H}_u) d\theta$, $\Delta_l^{(t)} := \int_0^1 (\mathbf{H}_l^{(t)}(\theta) - \mathbf{H}_l) d\theta$. However, it is still unclear from (4.12) how the iteration evolves around the strict saddle point.

To highlight ideas, let us define

$$\mathbf{M} := \mathbf{I} + \eta \mathbf{H}_l, \quad \mathbf{T} := \mathbf{I} - \eta \mathbf{H}_u. \tag{4.13}$$

It can be observed that $\mathbf{M}$ is a lower triangular matrix where the diagonal entries are all 1s; therefore it is invertible. After taking the inverse of matrix $\mathbf{M}$ on both sides of (4.12), we can obtain

$$\mathbf{x}^{(t+1)} = \mathbf{M}^{-1}\mathbf{T}\mathbf{x}^{(t)} - \eta \mathbf{M}^{-1}\Delta_u^{(t)}\mathbf{x}^{(t)} - \eta \mathbf{M}^{-1}\Delta_l^{(t)}\mathbf{x}^{(t+1)}.$$

Our goal of analyzing the recursion of $\mathbf{x}^{(t)}$ becomes to find the maximum eigenvalue of $\mathbf{M}^{-1}\mathbf{T}$. With the help of the matrix perturbation theory, we can quantify the difference between the eigenvalues of matrix $\mathbf{H}$ that contains the negative curvature and matrix $\mathbf{M}^{-1}\mathbf{T}$ that we are interested in analyzing. To be more precise, we give the following lemma.

Lemma 4. *Under Assumption 1, let* $\mathbf{H} := \nabla^2 f(\mathbf{x})$ *denote the Hessian matrix at an* ϵ*-SS2* $\mathbf{x}$ *where* $\lambda_{\min}(\mathbf{H}) \leq -\gamma$ *and* $\gamma > 0$*. We have*

$$\lambda_{\max}(\mathbf{M}^{-1}\mathbf{T}) > 1 + \frac{\eta\gamma}{1 + L/L_{\max}} \tag{4.14}$$

where $\mathbf{M}, \mathbf{T}$ *are defined in* (4.11) *and* (4.13)*.*

Lemma 4 illustrates that there exits a subspace spanned by the eigenvector of $\mathbf{M}^{-1}\mathbf{T}$ whose eigenvalue is greater than 1, indicating that the sequence generated by AGD can still potentially escape from the strict saddle point by leveraging such negative curvature information. Next, we can give a sketch of the proof of Theorem 8.

4.5.3 The Sketch of the Proof

The structure of the proof for quantifying the sufficient decrease of the objective function after the perturbation is borrowed from the proof of PGD (101), but PA-GD updates the variables block by block, so we have to provide the new proofs to show that PA-GD can still escape from saddle points with the perturbation technique.

First, if the size of the gradient is large enough, Algorithm 2 just implements the ordinary AGD. We give the descent lemma of AGD as follows.

Lemma 5. *Under Assumption 1, for the AGD algorithm with step size $\eta < 1/L_{\max}$, we have*

$$f(\mathbf{x}^{(t+1)}) \le f(\mathbf{x}^{(t)}) - \sum_{k=1}^{2} \frac{\eta}{2} \|\nabla_k f(\mathbf{h}_{-k}^{(t)}, \mathbf{x}_k^{(t)})\|^2.$$

Second, if the iterates are near a strict saddle point, we can show that the AGD algorithm after a perturbation can give a sufficient decrease with high probability in terms of the objective value. To be more precise, the statement is given as follows.

Lemma 6. *Under Assumption 1, there exists a absolute constant $c_{\max}$. Let $c \le c_{\max}$, $\chi \ge 1$, and η, r, g_{th}, t_{th} calculated as Algorithm 2 describes. Let $\widetilde{\mathbf{x}}^{(t)}$ be a strict saddle point, which satisfies*

$$\|\nabla f(\widetilde{\mathbf{x}}^{(t)})\|^2 \le 4 \sum_{k=1}^{2} \|\nabla_k f(\widetilde{\mathbf{h}}_{-k}^{(t)}, \widetilde{\mathbf{x}}_k^{(t)})\|^2 \le 4 g_{th}^2 \tag{4.15}$$

and $\lambda_{\min}(\nabla^2 f(\widetilde{\mathbf{x}}^{(t)})) \le -\gamma$, where $\widetilde{\mathbf{h}}_{-1}^{(t)} := \widetilde{\mathbf{x}}_2^{(t)}$ and $\widetilde{\mathbf{h}}_{-2}^{(t)} := \mathbf{x}_1^{(t+1)}$.

Let $\mathbf{x}^{(t)} = \widetilde{\mathbf{x}}^{(t)} + \xi^{(t)}$ where $\xi^{(t)}$ is generated randomly which follows the uniform distribution over $\mathbb{B}_0(r)$, and let $\mathbf{x}^{(t+t_{th})}$ be the iterates of PA-GD. With at least probability $1 - \frac{dL_{\max}}{(L_{\max}\rho\epsilon)^{1/3}} e^{-\chi}$, we have $f(\mathbf{x}^{(t+t_{th})}) - f(\widetilde{\mathbf{x}}^{(t)}) \le -f_{th}$.

We remark that Lemma 5 is well-known and Lemma 6 is the core technique. In the following, we outline the main idea used in proving the latter. The formal statements of these steps are shown in the appendix; see Lemma 14–Lemma 16 therein.

We emphasize that the main contributions of this work lies in the analysis of the first two steps, where the special update rule of PA-GD is analyzed so that the negative curvature of $\mathbf{H}$ around the saddle points can be utilized.

Step 1 (Lemma 14) Consider a generic sequence $\mathbf{u}^{(t)}$ generated by PA-GD. As long as the initial point of $\mathbf{u}^{(t)}$ is close to saddle point $\widetilde{\mathbf{x}}^{(t)}$, the distance between $\mathbf{u}^{(t)}$ and $\widetilde{\mathbf{x}}^{(t)}$ can be upper bounded by using the ρ-Hessian Lipschitz continuity property.

Step 2 (Lemma 15) Leveraging the negative curvature around the strict saddle point, we know that there exits a direction, i.e., $\vec{\mathbf{e}}$, which is spanned by the eigenvector of $\mathbf{M}^{-1}\mathbf{T}$ whose

corresponding eigenvalue is largest (greater than 1). Consider two sequences generated by PA-GD, $\mathbf{u}^{(t)}, \mathbf{w}^{(t)}$ initialized around the saddle point. When the initial points of these two iterates are separated apart away from each other along direction $\vec{\mathbf{e}}$ with a small distance, meaning that $\mathbf{w}^{(0)} = \mathbf{u}^{(0)} + \upsilon r\vec{\mathbf{e}},\ \upsilon \in [\delta/(2\sqrt{d}), 1]$ where r denotes the radius of the perturbation ball defined in Algorithm 2, we can show that if iterate $\mathbf{u}^{(t)}$ is still near the saddle point after T steps, the other sequence $\mathbf{w}^{(t)}$ will give a sufficient decrease of the objective value with less than T steps, implying that iterates $\mathbf{w}^{(t)}$ can escape from the saddle point with less than T steps.

Step 3 (Lemma 16) Consider $\mathbf{u}^{(0)}, \mathbf{w}^{(0)}$ as the points after the perturbation from the saddle point. We can quantify the probability that the AGD sequence will give a sufficient decrease of the objective value within T iterations after the perturbation (101, Lemma 14,15).

4.5.4 Extension to PA-PP

By leveraging the convergence analysis of PA-GD and relation between PA-GD and PA-PP shown in (4.7), we can also write the recursion of the PA-PP iteration as

$$\mathbf{x}^{(t+1)} + \eta\mathbf{H}'_l\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \eta\mathbf{H}'_u\mathbf{v}^{(t)} - \eta\Delta_u'^{(t)}\mathbf{x}^{(t)} - \eta\Delta_l'^{(t)}\mathbf{x}^{(t+1)} \tag{4.16}$$

where $\eta = 1/\nu$, $\Delta_u'^{(t)} := \int_0^1 (\mathbf{H}_u'^{(t)}(\theta) - \mathbf{H}'_u)d\theta$, $\Delta_l'^{(t)} := \int_0^1 (\mathbf{H}_l'^{(t)}(\theta) - \mathbf{H}'_l)d\theta$,

$$\mathbf{H}'_u = \begin{bmatrix} \mathbf{0} & \nabla^2_{12} f(\widetilde{\mathbf{x}}^{(t)}) \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \quad \mathbf{H}'_l = \begin{bmatrix} \nabla^2_{11} f(\widetilde{\mathbf{x}}^{(t)}) & \mathbf{0} \\ \nabla^2_{21} f(\widetilde{\mathbf{x}}^{(t)}) & \nabla^2_{22} f(\widetilde{\mathbf{x}}^{(t)}) \end{bmatrix}, \tag{4.17}$$

and

$$\mathbf{H}_l'^{(t)} := \begin{bmatrix} \nabla^2_{11} f(\theta\mathbf{x}_1^{(t+1)}, \theta\mathbf{x}_2^{(t)}) & \mathbf{0} \\ \\ \nabla^2_{21} f(\theta\mathbf{x}_1^{(t+1)}, \theta\mathbf{x}_2^{(t+1)}) & \nabla^2_{22} f(\theta\mathbf{x}_1^{(t+1)}, \theta\mathbf{x}_2^{(t+1)}) \end{bmatrix},$$

$$\mathbf{H}_u'^{(t)} := \begin{bmatrix} \mathbf{0} & \nabla^2_{12} f(\theta\mathbf{x}_1^{(t+1)}, \theta\mathbf{x}_2^{(t)}) \\ \\ \mathbf{0} & \mathbf{0} \end{bmatrix}. \tag{4.18}$$

Let

$$\mathbf{M}' := \mathbf{I} + \eta \mathbf{H}'_l \quad \mathbf{T}' := \mathbf{I} - \eta \mathbf{H}'_u. \tag{4.19}$$

We know that $\mathbf{T}'$ is an upper triangular matrix where the diagonal entries are all 1s, so it is invertible. Different from the case of PA-GD, we take the inverse of matrix $\mathbf{T}'$ on both sides of (4.16) and obtain

$$\mathbf{T}'^{-1}\mathbf{M}'\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \eta \mathbf{T}'^{-1}\Delta'^{(t)}_u \mathbf{x}^{(t)} - \eta \mathbf{T}'^{-1}\Delta'^{(t)}_l \mathbf{x}^{(t+1)}.$$

Then, we can give the following result that characterizes the recursion of $\mathbf{x}^{(t)}$ generated by PA-PP.

Corollary 4. *Under Assumption 1, let* $\mathbf{H} := \nabla^2 f(\mathbf{x})$ *denote the Hessian matrix at an* ϵ*-SS2* $\mathbf{x}$ *where* $\lambda_{\min}(\mathbf{H}) \leq -\gamma$ *and* $\gamma > 0$*. Let* $\lambda^+_{\min}(\cdot)$ *denote the minimum positive eigenvalue of a matrix. Then we have*

$$\lambda^+_{\min}(\mathbf{T}'^{-1}\mathbf{M}') \leq 1 - \eta\gamma/2 \tag{4.20}$$

where $\mathbf{M}', \mathbf{T}'$ *are defined in* (4.17) *and* (4.19)*;* $\eta \leq 1/L_{\max}$ *and* $\gamma \leq L_{\max}$*.*

We remark that Corollary 4 is useful since it can be leveraged to show that the norm of the iterates around saddle points can increase exponentially. Then, we can apply the similar analysis steps as the case of proving the convergence rate of PA-GD and obtain the results shown in Corollary 3.

4.6 Connection with Existing Works

Remark 3. In Theorem 8 we characterized the convergence rate to an $(\epsilon, \epsilon^{1/3})$-SS2. We can also translate this bound to the one for achieving an $(\epsilon, \sqrt{\epsilon})$-SS2, and in this case PA-GD needs $\widetilde{\mathcal{O}}(1/\epsilon^{3.5})$ iterations. Compared with the existing recent works (101), the convergence rate of PA-GD/PA-PP is slower than GD. The main reason is the fact that different from GD-type algorithms, PA-GD and PA-PP cannot fully utilize the Hessian information because they never see a full iteration. Similar situation happens for SGD-type of algorithms which also cannot get the exact negative curvature around strict saddle points.

From Table 4.1, it can be seen that the convergence rate of PA-GD/PA-PP is still faster than SGD (44), SGLD (92), Neon+SGD (94), and Neon2+SGD (95) to achieve an $(\epsilon, \sqrt{\epsilon})$-SS2, but slower than the rest. We emphasize that PA-GD and PA-PP represent the first BCD-type algorithms with the convergence rate guarantee to escape from the strict saddle points efficiently. At this point, it is unclear whether our rate is the best that is achievable, and the question of whether the resulting rate can be improved will be left to future work.

4.7 Numerical Results

4.7.1 A Simple Example

In this section, we present a simple example that shows the convergence behavior of PA-GD. Consider a nonconvex objective function, i.e.,

$$f(\mathbf{x}) := \mathbf{x}^T\mathbf{A}\mathbf{x} + \frac{1}{4}\|\mathbf{x}\|_4^4. \tag{4.21}$$

First, we have the following properties of function $f(\mathbf{x})$ such that $f(\mathbf{x})$ satisfies the assumptions of the analysis.

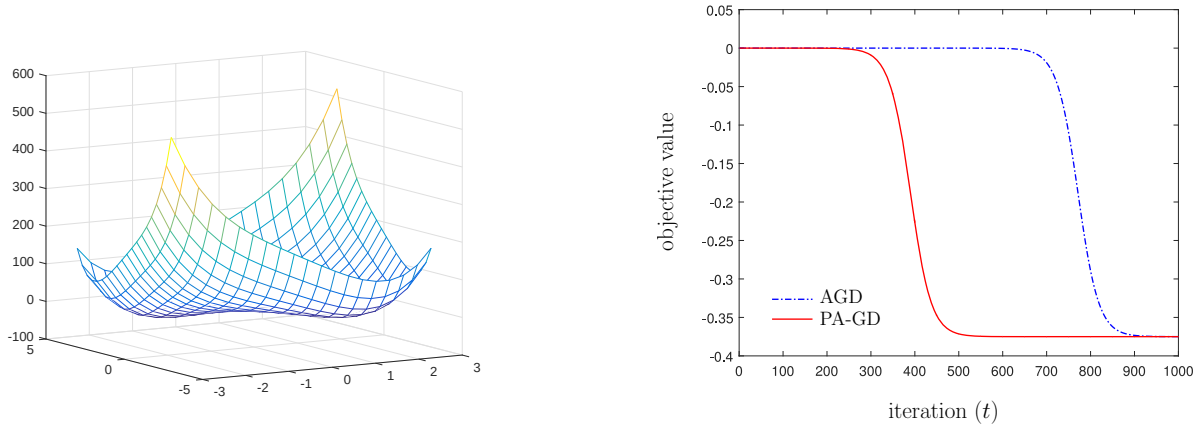
Lemma 7. *For any $\tau \geq \lambda_{\max}(\mathbf{A})$ and $\mathbf{x} \in \{\mathbf{x}|\|\mathbf{x}\|^2 \leq \tau\}$, $f(\mathbf{x})$ defined in (4.21) is 5τ-smooth and $6\sqrt{\tau}$-Hessian Lipschitz.*

Here, we can easily show the shape of objective function (4.21) in the two dimensional (2D) case in Figure 4.2(a), where $\mathbf{A} = [1\ 2; 2\ 1] \in \mathbb{R}^{2\times 2}$. It can be observed clearly that there exits a strict saddle point at $[0, 0]$ and two other local optimal points. We randomly initialize the algorithms around strict saddle point $[0, 0]$. The convergence comparison between AGD and PA-GD is shown in Figure 4.2(b). It can be observed that PA-GD converges faster than AGD to a local optimal point.

4.7.2 Asymmetric Matrix Factorization (AMF)

We consider a general asymmetric low rank matrix factorization problem as the following

$$\underset{\mathbf{U}\in\mathbb{R}^{n\times r},\mathbf{V}\in\mathbb{R}^{m\times r}}{\text{minimize}} \ \frac{1}{2}\|\mathbf{U}\mathbf{V}^T - \mathbf{M}\|_F^2. \tag{4.22}$$



(a) Objective function in 2D.

(b) Objective value versus the number of iterations

Figure 4.2 Convergence comparison between AGD and PA-GD, where $\epsilon = 10^{-4}$, $g_{\mathsf{th}} = \epsilon/10$, $\eta = 0.02$, $t_{\mathsf{th}} = 10/\epsilon^{1/3}$, $r = \epsilon/10$.

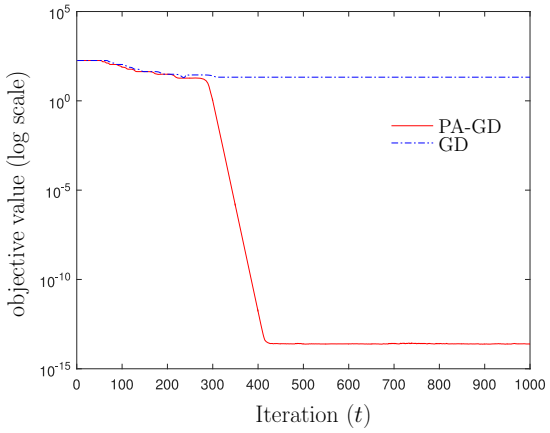


Figure 4.3 Convergence comparison between AGD and PA-GD for asymmetric matrix factorization, where $\epsilon = 10^{-14}$, $g_{\mathsf{th}} = \epsilon/10$, $\eta = 6 \times 10^{-3}$, $t_{\mathsf{th}} = 10/\epsilon^{1/3}$, $r = \epsilon/10$.

However, the global optimal solution has a scaling ambiguity problem (106). In (106), it is shown that a reformulated problem of (4.22) is

$$\underset{\mathbf{U}\in\mathbb{R}^{n\times r},\mathbf{V}\in\mathbb{R}^{m\times r}}{\text{minimize}} \; g(\mathbf{W}) = f(\mathbf{W}) + \rho(\mathbf{W}) \tag{4.23}$$

where

$$f(\mathbf{W}) := \frac{1}{2}\|\mathbf{U}\mathbf{V}^T - \mathbf{M}\|_F^2, \quad \rho(\mathbf{W}) := \frac{\mu}{4}\|\mathbf{U}^T\mathbf{U} - \mathbf{V}^T\mathbf{V}\|_F^2, \quad \mathbf{W} := \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix}, \quad \mu > 0.$$

This problem has the same global optimal solution as (4.22). Also, all saddle points of the problems are strict and within a ball with certain radius, and every local optimal points of this problem is global optimal (106, Theorem 1). Therefore, as long as the algorithm can escape from the saddle points, the algorithm will converge to the global optimal solution.

In the simulation results, we randomly generate matrix $\mathbf{M} = \mathbf{U}\mathbf{V}^T$ with dimension $n = 200, m = 20, r = 10$ and initialize GD and PA-GD around saddle point 0. GD and PA-GD use the same step size, which is η shown in Figure 4.3. It can be observed that PA-GD can escape the saddle point much faster than GD and converge to the global optimal solution.

CHAPTER 5. CONCLUSION

In this dissertation, the first-order methods of solving nonconvex optimization problems are studied for both constrained and unconstrained cases. The main work focuses on the nonconvex algorithm design, convergence analysis, and optimality analysis of the obtained solutions. The principle applications of the algorithms are matrix factorization related topics, such as SymNMF, stochastic SymNMF, AMF, etc.

In the constrained nonoconvex optimization problems, we propose a nonconvex splitting algorithm for solving the SymNMF problem. We show that the proposed algorithm converges to a KKT point in a sublinear manner. Further, we provide sufficient conditions to identify global or local optimal solutions of the SymNMF problem. Numerical experiments show that the proposed method can converge quickly to local optimal solutions.

The stochastic SymNMF problem is considered in the areas of clustering and community detection. We show that the proposed stochastic nonconvex splitting algorithm converges to the set of stationary points of SymNMF in a sublinear manner. Numerical experiments show that the proposed method has a similar convergence rate and clustering accuracy as deterministic SymNMF does.

In the future, we plan to extend the proposed methods in a way such that the algorithms can converge to the local or even global optimal solutions of SymNMF without requiring checking conditions. Also, it is possible to apply the nonconvex splitting method to more general matrix factorization problems, such as the quadratic nonnegative matrix factorization problem.

The perturbed variants of AGD and alternating proximal point (APP) algorithms are proposed, with the objective of finding the second order stationary solutions of nonconvex smooth problems. Leveraging the recently developed idea of random perturbation for the first-order methods, the proposed algorithms add suitable perturbation to the AGD or APP iterates. The main contribution

of this work is a new analysis that takes into consideration the block structure of the updates for the perturbed AGD and APP algorithms. By exploiting the negative curvature, it is established that with high probability the algorithms can converge to an $(\epsilon, \epsilon^{1/3})$-SS2 with $\mathcal{O}(\text{polylog}(d)/\epsilon^{7/3})$ iterations.

BIBLIOGRAPHY

[1] S. Campbell and G. Poole, "Computing nonnegative rank factorizations," *Linear Algebra and its Applications*, vol. 35, pp. 175–182, Feb. 1981.

[2] P. Paatero and U. Tapper, "Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values," *Environmetrics*, vol. 5, no. 2, pp. 111–126, June 1994.

[3] N. Gillis and S. A. Vavasis, "Fast and robust recursive algorithmsfor separable nonnegative matrix factorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 4, pp. 698–714, Apr. 2014.

[4] Y. Wang and Y. Zhang, "Nonnegative matrix factorization: A comprehensive review," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 6, pp. 1336–1353, June 2013.

[5] P. Hoyer, "Non-negative matrix factorization with sparseness constraints," *Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, 2004.

[6] D. Lee and H. Seung, "Algorithms for non-negative matrix factorization," in *Proceedings of Neural Information Processing Systems (NIPS)*, 2001.

[7] B. Yang, X. Fu, and N. D. Sidiropoulos, "Joint factor analysis and latent clustering," in *Proceedings of IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, Dec. 2015, pp. 173–176.

[8] N. Gillis, "The why and how of nonnegative matrix factorization," in *Regularization, Optimization, Kernels, and Support Vector Machines.* Chapman & Hall/CRC, Machine Learning and Pattern Recognition Series, 2014.

[9] Z. He, S. Xie, R. Zdunek, G. Zhou, and A. Cichocki, "Symmetric nonnegative matrix factorization: Algorithms and applications to probabilistic clustering," *IEEE Transactions on Neural Networks*, vol. 22, no. 12, pp. 2117–2131, Dec. 2011.

[10] K. Huang, N. Sidiropoulos, and A. Swami, "Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition," *IEEE Transactions on Signal Processing*, vol. 62, no. 1, pp. 211–224, Jan. 2014.

[11] D. Kuang, S. Yun, and H. Park, "SymNMF: nonnegative low-rank approximation of a similarity matrix for graph clustering," *Journal of Global Optimization*, vol. 62, no. 3, pp. 545–574, July 2015.

[12] F. Wang, T. Li, X. Wang, S. Zhu, and C. Ding, "Community discovery using nonnegative matrix factorization," *Data Mining and Knowledge Discovery*, vol. 22, no. 3, pp. 493–521, May 2011.

[13] T. Gao, S. Olofsson, and S. Lu, "Minimum-volume-regularized weighted symmetric nonnegative matrix factorization for clustering," in *Proceedings of IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Dec. 2016, pp. 247–251.

[14] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, 2007.

[15] S. Lu and Z. Wang, "Accelerated algorithms for eigen-value decomposition with application to spectral clustering," in *Proceedings of Asilomar Conference on Signals, Systems and Computers (Asilomar)*, Nov. 2015, pp. 355–359.

[16] C. Ding, X. He, and H. Simon, "On the equivalence of nonnegative matrix factorization and spectral clustering." in *Proceedings of SIAM International Conference on Data Mining*, vol. 5, 2005, pp. 606–610.

[17] S. Lu, M. Hong, and Z. Wang, "A nonconvex splitting method for symmetric nonnegative matrix factorization: Convergence analysis and optimality," in *Proceedings of IEEE Inter-*

national Conference on Acoustics Speech and Signal Process (ICASSP), March 2017, pp. 2572–2576.

[18] D. Sussman, M. Tang, D. Fishkind, and C. Priebe, "A consistent adjacency spectral embedding for stochastic blockmodel graphs," *Journal of the American Statistical Association*, vol. 107, no. 499, pp. 1119–1128, 2012.

[19] P. Jain, P. Netrapalli, and S. Sanghavi, "Low-rank matrix completion using alternating minimization," in *Proc. of the Forty-Fifth Annual ACM Symposium on Theory of Computing*, 2013, pp. 665–674.

[20] L. Balzano, R. Nowak, and B. Recht, "Online identification and tracking of subspaces from highly incomplete information," in *Proc. of Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2010, pp. 704–711.

[21] E. Abbe and C. Sandon, "Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery," in *Proceedings of IEEE the 56th Annual Symposium on Foundations of Computer Science (FOCS)*, 2015, pp. 670–688.

[22] A. Bifet, G. Holmes, R. Kirkby, and B. Pfahringer, "MOA: Massive online analysis," *Journal of Machine Learning Research*, vol. 11, no. May, pp. 1601–1604, 2010.

[23] Y. Chen and Y. Chi, "Harnessing structures in big data via guaranteed low-rank matrix estimation," *IEEE Signal Processing Magazine*, 2018.

[24] R. Ge, C. Jin, and Y. Zheng, "No spurious local minima in nonconvex low rank problems: A unified geometric analysis," in *Proceedings of International Conference on Machine Learning (ICML)*, 2017, pp. 1233–1242.

[25] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, ser. Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2001.

[26] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

[27] D. P. Bertsekas, *Nonlinear Programming, 2nd.* Belmont, MA: Athena Scientific, 1999.

[28] Y. Li and Y. Liang, "Provable alternating gradient descent for non-negative matrix factorization with strong correlations," in *Proceedings of International Conference on Machine Learning (ICML)*, vol. 70, 2017, pp. 2062–2070.

[29] Y. Nesterov, "Efficiency of coordinate descent methods on huge-scale optimization problems," *SIAM Journal on Optimization*, vol. 22, no. 2, pp. 341–362, 2012.

[30] A. Beck and L. Tetruashvili, "On the convergence of block coordinate descent type methods," *SIAM Journal on Optimization*, vol. 23, no. 4, pp. 2037–2060, 2013.

[31] C. Mai, S. Lu, J. Sun, and G. Wang, "Beampattern optimization for frequency diverse array with sparse frequency waveforms," *IEEE Access*, vol. 5, pp. 17 914–17 926, 2017.

[32] M. Razaviyayn, M. Hong, and Z. Luo, "A unified convergence analysis of block successive minimization methods for nonsmooth optimization," *SIAM Journal on Optimization*, vol. 23, no. 2, pp. 1126–1153, 2013.

[33] M. Hong, X. Wang, M. Razaviyayn, and Z. Luo, "Iteration complexity analysis of block coordinate descent methods," *Mathematical Programming Series A*, vol. 163, no. 1, pp. 85–114, May 2017.

[34] P. Tseng, "Convergence of a block coordinate descent method for nondifferentiable minimization," *Journal of Optimization Theory and Applications*, vol. 109, no. 3, pp. 475–494, 2001.

[35] P. Tseng and S. Yun, "Block-coordinate gradient descent method for linearly constrained nonsmooth separable optimization," *Journal of Optimization Theory and Applications*, vol. 140, no. 3, p. 513, 2009.

[36] T. Zhao, Z. Wang, and H. Liu, "A nonconvex optimization framework for low rank matrix estimation," in *Proceedings of Neural Information Processing Systems (NIPS)*, 2015, pp. 559–567.

[37] Y. Xu and W. Yin, "A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion," *SIAM Journal on Imaging Sciences*, vol. 6, no. 3, pp. 1758–1789, 2013.

[38] Z. Zhang and M. Brand, "On the convergence of block coordinate descent in training DNNs with Tikhonov regularization," in *Proceedings of Neural Information Processing Systems (NIPS)*, 2017.

[39] L. Grippo and M. Sciandrone, "On the convergence of the block nonlinear Gauss-Seidel method under convex constraints," *Operations Research Letters*, vol. 26, pp. 127–136, 2000.

[40] Q. Shi, H. Sun, S. Lu, M. Hong, and M. Razaviyayn, "Inexact block coordinate descent methods for symmetric nonnegative matrix factorization," *IEEE Transactions on Signal Processing*, vol. 65, no. 22, pp. 5995–6008, Nov. 2017.

[41] M. Razaviyayn, M. Hong, Z. Luo, and J. Pang, "Parallel successive convex approximation for nonsmooth nonconvex optimization," in *Proceedings of Neural Information Processing Systems (NIPS)*, 2014.

[42] K. Kawaguchi, "Deep learning without poor local minima," in *Proceedings of Neural Information Processing Systems (NIPS)*, 2016, pp. 586–594.

[43] S. Feizi, H. Javadi, J. Zhang, and D. Tse, "Porcupine neural networks: (almost) all local optima are global," *arXiv:1710.02196 [stat.ML]*, 2017.

[44] R. Ge, F. Huang, C. Jin, and Y. Yuan, "Escaping from saddle points — online stochastic gradient for tensor decomposition," in *Proceedings of Annual Conference on Learning Theory (COLT)*, 2015, pp. 797–842.

[45] J. Sun, Q. Qu, and J. Wright, "When are nonconvex problems not scary?" in *Proceedings of NIPS Workshop on Non-convex Optimization for Machine Learning: Theory and Practice*, 2015.

[46] J. Sun, Q. Qu, and J. Wright, "A geometric analysis of phase retrieval," *arXiv:1602.06664 [cs.IT]*, 2017.

[47] G. Wang, G. Giannakis, Y. Saad, and J. Chen, "Solving almost all systems of random quadratic equations," in *Proceedings of Neural Information Processing Systems (NIPS)*, 2017.

[48] C. Lin, "Projected gradient methods for nonnegative matrix factorization," *Neural computation*, vol. 19, no. 10, pp. 2756–2779, 2007.

[49] J. Kim and H. Park, "Fast nonnegative matrix factorization: An active-set-like method and comparisons," *SIAM Journal on Scientific Computing*, vol. 33, no. 6, pp. 3261–3281, 2011.

[50] J. Parker, P. Schniter, and V. Cevher, "Bilinear generalized approximate message passing – part i: Derivation," *IEEE Transactions on Signal Processing*, vol. 62, no. 22, pp. 5839–5853, Nov. 2014.

[51] J. Parker, P. Schniter, and V. Cevher, "Bilinear generalized approximate message passing – part ii: Applications," *IEEE Transactions on Signal Processing*, vol. 62, no. 22, pp. 5854–5867, Nov. 2014.

[52] J. Kim, Y. He, and H. Park, "Algorithms for nonnegative matrix and tensor factorizations: A unified view based on block coordinate descent framework," *Journal of Global Optimization*, vol. 58, no. 2, pp. 285–319, Mar. 2013.

[53] C. Lin, "On the convergence of multiplicative update algorithms for nonnegative matrix factorization," *IEEE Transactions on Neural Networks*, vol. 18, no. 6, pp. 1589–1596, Nov. 2007.

[54] D. Kuang, C. Ding, and H. Park, "Symmetric nonnegative matrix factorization for graph clustering," in *Proceedings of SIAM International Conference on Data Mining*, 2012, pp. 106–117.

[55] A. Vandaele, N. Gillis, Q. Lei, K. Zhong, and I. Dhillon, "Efficient and non-convex coordinate descent for symmetric nonnegative matrix factorization," *IEEE Transactions on Signal Processing*, vol. 64, no. 21, pp. 5571–5584, Nov. 2016.

[56] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[57] Y. Xu, W. Yin, Z. Wen, and Y. Zhang, "An alternating direction algorithm for matrix completion with non-negative factors," *Frontiers of Mathematics in China*, vol. 7, no. 2, pp. 365–384, June 2012.

[58] D. Sun and C. Fevotte, "Alternating direction method of multipliers for non-negative matrix factorization with the beta-divergence," in *Proceedings of IEEE International Conference on Acoustics Speech and Signal Process (ICASSP)*, May 2014, pp. 6201–6205.

[59] K. Huang, N. D. Sidiropoulos, and A. P. Liavas, "A flexible and effcient algorithmic framework for constrained matrix and tensor factorization," *IEEE Transactions on Signal Processing*, vol. 64, no. 19, pp. 5052–5065, June 2016.

[60] S. Vavasis, "On the complexity of nonnegative matrix factorization," *SIAM Journal on Optimization*, vol. 20, no. 3, pp. 1364–1377, 2009.

[61] P. Dickinson and L. Gijben, "On the computational complexity of membership problems for the completely positive cone and its dual," *Computational Optimization and Applications*, vol. 57, pp. 403–415, Mar. 2014.

[62] C. Sa, C. Re, and K. Olukotun, "Global convergence of stochastic gradient descent for some non-convex matrix problems," in *Proceedings of International Conference on Machine Learning (ICML)*, 2015, pp. 2332–2341.

[63] N. Gillis, "Sparse and unique nonnegative matrix factorization through data preprocessing," *Journal of Machine Learning Research*, vol. 13, pp. 3349–3386, 2012.

[64] R. Sun and Z.-Q. Luo, "Guaranteed matrix completion via non-convex factorization," *IEEE Transactions on Information Theory*, vol. 62, no. 11, pp. 6535–6579, Nov. 2016.

[65] A. Montanari and E. Richard, "Non-negative principal component analysis:message passing algorithms and sharp asymptotics," *IEEE Transactions on Information Theory*, vol. 62, no. 3, pp. 1458–1484, 2016.

[66] D. P. Bertsekas, P. Hosein, and P. Tseng, "Relaxation methods for network flow problems with convex arc costs," *SIAM Journal on Control and Optimization*, vol. 25, no. 5, pp. 1219–1243, Sept. 1987.

[67] J. Eckstein and D. P. Bertsekas, "On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators," *Mathematical Programming*, vol. 55, no. 1, pp. 293–318, 1992.

[68] M. Hong, Z.-Q. Luo, and M. Razaviyayn, "Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems," *SIAM Journal on Optimization*, vol. 26, no. 1, pp. 337–364, 2016.

[69] G. Li and T.-K. Pong, "Global convergence of splitting methods for nonconvex composite optimization," *SIAM Journal on Optimization*, vol. 25, no. 4, pp. 2434–2460, 2015.

[70] B. Ames and M. Hong, "Alternating direction method of multipliers for penalized zero-variance discriminant analysis," *Computational Optimization and Applications*, vol. 64, no. 3, pp. 725–754, 2016.

[71] Y. Wang and J. Z. W. Yin, "Global convergence of ADMM in nonconvex nonsmooth optimization," *arXiv Preprint, arXiv:1511.06324*, 2015.

[72] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[73] F. Facchinei and J. Pang, *Finite-dimensional variational inequalities and complementarity problems*. Springer Science & Business Media, 2007.

[74] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imgaging Science*, vol. 2, no. 1, pp. 183–202, 2009.

[75] M. Razaviyayn, M. Hong, Z. Luo, and J. Pang, "Parallel successive convex approximation for nonsmooth nonconvex optimization," in *Proceedings of Neural Information Processing Systems (NIPS)*, 2014, pp. 1440–1448.

[76] G. Scutari, F. Facchinei, P. Song, D. P. Palomar, and J. S. Pang, "Decomposition by partial linearization: Parallel optimization of multi-agent systems," *IEEE Transactions on Signal Processing*, vol. 62, no. 3, pp. 641–656, Feb. 2014.

[77] C. Navasca, L. De Lathauwer, and S. Kindermann, "Swamp reducing technique for tensor decomposition," in *Proceedings of the 16th European Signal Processing Conference*, 2008, pp. 1–5.

[78] D. Cai, X. He, and J. Han, "Locally consistent concept factorization for document clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 6, pp. 902–913, 2011.

[79] J. Leskovec, K. Lang, A. Dasgupta, and M. Mahoney, "Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters." *Internet Mathematics*, vol. 6, no. 1, pp. 29–123, 2009.

[80] E. Cho, S. Myers, and J. Leskovec, "Friendship and mobility: user movement in location-based social networks," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2011, pp. 1082–1090.

[81] M. Razaviyayn, M. Sanjabi, and Z. Luo, "A stochastic successive minimization method for nonsmooth nonconvex optimization with applications to transceiver design in wireless communication networks," *Mathematical Programming*, vol. 157, no. 2, pp. 515–545, 2016.

[82] S. Ghadimi, G. Lan, and H. Zhang, "Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization," *Mathematical Programming*, vol. 155, no. 1-2, pp. 267–305, 2016.

[83] H. Ouyang, N. He, L. Tran, and A. Gray, "Stochastic alternating direction method of multipliers," in *Proceedings of the 30th International Conference on Machine Learning*, 2013, pp. 80–88.

[84] S. Azadi and S. Sra, "Towards an optimal stochastic alternating direction method of multipliers," in *Proceedings of the 31st International Conference on Machine Learning*, 2014, pp. 620–628.

[85] S. Lu, M. Hong, and Z. Wang, "A nonconvex splitting method for symmetric nonnegative matrix factorization: Convergence analysis and optimality," *IEEE Transactions on Signal Processing*, vol. 65, no. 12, pp. 3120–3135, June 2017.

[86] S. Lu, M. Hong, and Z. Wang, "A stochastic nonconvex splitting method for symmetric nonnegative matrix factorization," in *Proceedings of International Conference on Artificial Intelligence and Statistics (AISTATS)*, vol. 54, 2017, pp. 812–821.

[87] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering," in *Proceedings of Neural Information Processing Systems (NIPS)*, 2004.

[88] A. Conn, N. Gould, and P. Toint, *Trust region methods*. SIAM, 2000.

[89] Y. Nesterov and B. Polyak, "Cubic regularization of Newton method and its global performance," *Mathematical Programming*, vol. 108, no. 1, pp. 177–205, 2006.

[90] Y. Carmon and J. Duchi, "Gradient descent efficiently finds the cubic-regularized non-convex Newton step," *arXiv preprint arXiv:1612.00547*, 2016.

[91] S. Reddi, M. Zaheer, S. Sra, B. Póczos, F. Bach, R. Salakhutdinov, and A. Smola, "A generic approach for escaping saddle points," in *Proceedings of International Conference on Artificial Intelligence and Statistics (AISTATS)*, vol. 84, 2018, pp. 1233–1242.

[92] Y. Zhang, P. Liang, and M. Charikar, "A hitting time analysis of stochastic gradient langevin dynamics," in *Proceedings of Annual Conference on Learning Theory (COLT)*, 2017, pp. 1980–2022.

[93] M. Raginsky, A. Rakhlin, and M. Telgarsky, "Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis," in *Proceedings of Annual Conference on Learning Theory (COLT)*, 2017, pp. 1674–1703.

[94] Y. Xu and T. Yang, "First-order stochastic algorithms for escaping from saddle points in almost linear time," *arXiv preprint arXiv:1711.01944*, 2017.

[95] Z. Allen-Zhu and Y. Li, "Neon2: Finding local minima via first-order oracles," *arXiv preprint arXiv:1711.06673*, 2017.

[96] Y. Xu, R. Jin, and T. Yang, "Neon+: Accelerated gradient methods for extracting negative curvature for non-convex optimization," *arXiv preprint arXiv:1712.01033*, 2017.

[97] Y. Yu, D. Zou, and Q. Gu, "Saving gradient and negative curvature computations: Finding local minima more efficiently," *arXiv preprint arXiv:1712.03950*, 2017.

[98] J. Lee, M. Simchowitz, M. Jordan, and B. Recht, "Gradient descent only converges to minimizers," in *Proceedings of Annual Conference on Learning Theory (COLT)*, 2016, pp. 1246–1257.

[99] J. Lee, I. Panageas, G. Piliouras, M. Simchowitz, M. Jordan, and B. Recht, "First-order methods almost always avoid saddle points," *arXiv:1710.07406v1 [stat.ML]*, 2017.

[100] S. Du, C. Jin, J. Lee, M. Jordan, B. Póczos, and A. Singh, "Gradient descent can take exponential time to escape saddle points," in *Proceedings of Neural Information Processing Systems (NIPS)*, 2017, pp. 1067–1077.

[101] C. Jin, R. Ge, P. Netrapalli, S. Kakade, and M. Jordan, "How to escape saddle points efficiently," in *Proceedings of International Conference on Machine Learning (ICML)*, 2017, pp. 1724–1732.

[102] C. Jin, P. Netrapalli, and M. Jordan, "Accelerated gradient descent escapes saddle points faster than gradient descent," in *Proceedings of Annual Conference on Learning Theory (COLT)*, vol. 75, 2018, pp. 1042–1085.

[103] S. Lu, M. Hong, and Z. Wang, "On the sublinear convergence of randomly perturbed alternating gradient descent to second order stationary solutions," *arXiv preprint arXiv:1802.10418*, 2018.

[104] R. Sun and M. Hong, "Improved iteration complexity bounds of cyclic block coordinate descent for convex problems," in *Proceedings of Neural Information Processing Systems (NIPS)*, 2015, pp. 1306–1314.

[105] N. Parikh and S. Boyd, "Proximal algorithms," *Foundations and Trends in Optimization*, vol. 1, no. 3, pp. 127–239, 2014.

[106] Z. Zhu, Q. Li, G. Tang, and M. Wakin, "Global optimality in low-rank matrix optimization," *arXiv:1702.07945*, 2017.

[107] H. Weyl, "Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung)," *Mathematische Annalen*, vol. 71, no. 4, pp. 441–479, 1912.

[108] J. Holbrook, "Spectral variation of normal matrices," *Linear Algebra and its Applications*, vol. 174, pp. 131–144, 1992.

[109] J. Angelos, C. Cowen, and S. Narayan, "Triangular truncation and finding the norm of a Hadamard multiplier," *Linear Algebra and its Applications*, vol. 170, pp. 117–135, 1992.

APPENDIX A. SOME PROOFS OF SYMNMF

A.1 Proof of Lemma 1

Sufficiency: the stationary points satisfy

$$\left\langle \big(\mathbf{X}^*(\mathbf{X}^*)^T - (\mathbf{Z}^T + \mathbf{Z})/2\big)\mathbf{X}^*, \mathbf{X} - \mathbf{X}^* \right\rangle \geq 0, \quad \forall\, \mathbf{X} \geq 0. \tag{A.1}$$

Let $\mathbf{\Omega} \triangleq (\mathbf{X}^*(\mathbf{X}^*)^T - (\mathbf{Z}^T + \mathbf{Z})/2)\mathbf{X}^*/2$. We have $\langle \mathbf{\Omega}, \mathbf{X} - \mathbf{X}^* \rangle \geq 0, \forall \mathbf{X} \geq 0$. By setting $\mathbf{X}$ appropriately as $0 \leq \mathbf{X} \leq \mathbf{X}^*$, we have $\mathbf{\Omega}_{i,j} \geq 0, (i,j) \in \mathcal{S}$ where $\mathcal{S} = \{i, j | \mathbf{X}^*_{i,j} \neq 0\}$. Also, by setting $\mathbf{X}$ appropriately as $\mathbf{X} \geq \mathbf{X}^*$, we have $\mathbf{\Omega}_{i,j} \geq 0, (i,j) \notin \mathcal{S}$. Combining the two cases, we conclude that $\mathbf{\Omega} \geq 0$.

From (A.1), we know that $\langle \mathbf{\Omega}, \mathbf{X} \rangle \geq \langle \mathbf{\Omega}, \mathbf{X}^* \rangle$. Since $\mathbf{\Omega} \geq 0$ and $\mathbf{X} \geq 0$, we have $\langle \mathbf{\Omega}, \mathbf{X} \rangle \geq 0, \forall \mathbf{X}$, meaning that $\langle \mathbf{\Omega}, \mathbf{X}^* \rangle \leq 0$. Combining with $\mathbf{X}^* \geq 0$ and $\mathbf{\Omega} \geq 0$, we have $\langle \mathbf{\Omega}, \mathbf{X}^* \rangle \geq 0$, which results in $\langle \mathbf{\Omega}, \mathbf{X}^* \rangle = 0$.

In summary, we have

$$2\left(\mathbf{X}^*(\mathbf{X}^*)^T - \frac{\mathbf{Z}^T + \mathbf{Z}}{2}\right)\mathbf{X}^* - \mathbf{\Omega} = 0, \tag{A.2a}$$

$$\mathbf{\Omega} \geq 0, \tag{A.2b}$$

$$\mathbf{X}^* \geq 0, \tag{A.2c}$$

$$\langle \mathbf{X}^*, \mathbf{\Omega} \rangle = 0, \tag{A.2d}$$

which are the KKT conditions of the SymNMF problem.

Necessity: If the point is a KKT point of SymNMF, we have

$$\mathbf{\Omega}^* = 2\big(\mathbf{X}^*(\mathbf{X}^*)^T - \frac{\mathbf{Z}^{T+}\mathbf{Z}}{2}\big)\mathbf{X}^*. \tag{A.3}$$

Combining with $\langle \mathbf{X}^*, \mathbf{\Omega} \rangle = 0$, we know that

$$\langle \mathbf{\Omega}^*, \mathbf{X} - \mathbf{X}^* \rangle \geq 0, \quad \forall\, \mathbf{X} \geq 0, \tag{A.4}$$

which is the condition of stationary points.

A.2 Proof of Lemma 2

In this section, we prove the equivalence between KKT points of (1.1) and those of (2.3).

Proof: Below we show that if τ is large enough, then the KKT conditions of (1.1) and (2.3) are the same. It is sufficient to show that when τ is large enough, there can be no KKT point whose column has size τ, leading to the fact that the constraint $\|\mathbf{X}_k^*\|^2 \leq \tau$ is always inactive.

We check the optimality condition of the SymNMF problem at $\|\mathbf{X}_k^*\|^2 = \tau_k$, where $\tau_k > 0$ is a constant. We can rewrite the objective function as

$$\begin{aligned} f(\mathbf{X}) =& \frac{1}{2}\Bigg(\sum_{i=1,i\neq k}^{N} \sum_{j=1,j\neq k}^{N} (\mathbf{X}_i\mathbf{X}_j^T - \mathbf{Z}_{i,j})^2 + \sum_{i=1,i\neq k}^{N} (\mathbf{X}_i\mathbf{X}_k^T - \mathbf{Z}_{i,k})^2 \\ &+ \sum_{j=1,j\neq k}^{N} (\mathbf{X}_k\mathbf{X}_j^T - \mathbf{Z}_{k,j})^2 + (\mathbf{X}_k\mathbf{X}_k^T - \mathbf{Z}_{k,k})^2 \Bigg). \end{aligned}$$

Note, $\mathbf{X}_i, \mathbf{X}_j, \mathbf{X}_k$ denote rows of the matrix $\mathbf{X}$.

We take the gradient of $f(\mathbf{X})$ with respective to $\mathbf{X}_k$ and obtain

$$\begin{aligned} \frac{\partial f(\mathbf{X})}{\partial \mathbf{X}_{k,m}} &= \sum_{i=1,i\neq k}^{N} \mathbf{X}_{i,m}(\mathbf{X}_i\mathbf{X}_k^T - \mathbf{Z}_{i,k}) \sum_{j=1,j\neq k}^{N} \mathbf{X}_{j,m}(\mathbf{X}_k\mathbf{X}_j^T - \mathbf{Z}_{k,j}) + 2\mathbf{X}_{k,m}(\mathbf{X}_k\mathbf{X}_k^T - \mathbf{Z}_{k,k}) \\ &= \sum_{i=1,i\neq k}^{N} \mathbf{X}_{i,m}(\mathbf{X}_i\mathbf{X}_k^T - (\mathbf{Z}_{i,k} + \mathbf{Z}_{k,i})) + 2\mathbf{X}_{k,m}(\mathbf{X}_k\mathbf{X}_k^T - \mathbf{Z}_{k,k}) \end{aligned} \tag{A.5}$$

where $\mathbf{X}_{i,m}$ denotes the mth entry of the ith row of $\mathbf{X}$.

Assume that $\mathbf{X}_k^*$ is a KKT point. We have $(\frac{\partial f(\mathbf{X}_k^*)}{\partial \mathbf{X}_k})(\mathbf{X}_k - \mathbf{X}_k^*)^T \geq 0, \forall\ \mathbf{X}_k \in \mathcal{X}$, where $\mathcal{X} = \{\mathbf{X}_k | \mathbf{X}_k \geq 0, \|\mathbf{X}_k\|^2 \leq \tau_k\}$, which implies

$$\begin{aligned} &\frac{\partial f(\mathbf{X}_k^*)}{\partial \mathbf{X}_{k,m}}(\mathbf{X}_{k,m} - \mathbf{X}_{k,m}^*) \geq 0 \\ &0 \leq \mathbf{X}_{k,m} \leq \mathbf{X}_{k,m}^* = \sqrt{\tau_k - \sum_{n=1,n\neq m}^{K} (\mathbf{X}_{k,n}^*)^2} \quad \forall\ m. \end{aligned} \tag{A.6}$$

Since $\|\mathbf{X}_k^*\|^2 = \tau_k$, then there exists an index m such that $\mathbf{X}_{k,m}^* > 0$. Consider a feasible point $0 \leq \mathbf{X}_{k,m} < \mathbf{X}_{k,m}^*$, where $m \in \mathcal{S}_m \triangleq \{m | \mathbf{X}_{k,m}^* \neq 0\}$. According to (A.6), we have

$$\frac{\partial f(\mathbf{X}_{k,m}^*)}{\partial \mathbf{X}_{k,m}} \leq 0, \quad 0 \leq \mathbf{X}_{k,m} < \mathbf{X}_{k,m}^* \quad \forall\ m \in \mathcal{S}_m. \tag{A.7}$$

Plugging (A.5) into (A.7) and multiplying $\mathbf{X}^*_{k,m}$ on both sides of (A.7), we can obtain

$$\mathbf{X}^*_{k,m}\Big(\sum_{i=1,i\neq k}^{N}\mathbf{X}^*_{i,m}\big(\mathbf{X}^*_i(\mathbf{X}^*_k)^T-\frac{\mathbf{Z}_{i,k}+\mathbf{Z}_{k,i}}{2}\big)+\mathbf{X}^*_{k,m}(\mathbf{X}^*_k(\mathbf{X}^*_k)^T-\mathbf{Z}_{k,k})\Big)\leq 0\quad\forall\ m\in\mathcal{S}_m. \tag{A.8}$$

For the case $m\notin\mathcal{S}_m$, we know that $\mathbf{X}^*_{k,m}=0$. Summing up (A.8) $\forall m$, and noting that $|\mathcal{S}_m|\geq 1$ we can get

$$p\triangleq\sum_{i=1,i\neq k}^{N}\underbrace{\mathbf{X}^*_i(\mathbf{X}^*_k)^T\big(\mathbf{X}^*_i(\mathbf{X}^*_k)^T-\frac{\mathbf{Z}_{i,k}+\mathbf{Z}_{k,i}}{2}\big)}_{\triangleq\mathcal{M}_{i,k}}+\mathbf{X}^*_k(\mathbf{X}^*_k)^T(\mathbf{X}^*_k(\mathbf{X}^*_k)^T-\mathbf{Z}_{k,k})\leq 0. \tag{A.9}$$

In (A.9), $\mathcal{M}_{i,k}$ is a quadratic function with respective to $C_{i,k}$, where $C_{i,k}\triangleq\mathbf{X}^*_i(\mathbf{X}^*_k)^T$, so the minimum of $\mathcal{M}_{i,k}$ is $-1/4((\mathbf{Z}_{i,k}+\mathbf{Z}_{k,i})/2)^2$. Consequently, the minimum of $\sum_{i=1,i\neq k}^{N}\mathcal{M}_{i,k}$ is $-1/4\sum_{i=1,i\neq k}^{N}((\mathbf{Z}_{i,k}+\mathbf{Z}_{k,i})/2)^2$.

In addition, since we have $\|\mathbf{X}^*_k\|^2=\tau_k$, the lower bound of p is $p_{\text{L}}\triangleq-1/4\sum_{i=1,i\neq k}^{N}((\mathbf{Z}_{i,k}+\mathbf{Z}_{k,i})/2)^2+\tau_k(\tau_k-\mathbf{Z}_{k,k})$ which is a quadratic function in terms of τ_k. Therefore, we have that if

$$\tau_k>\theta_k\triangleq\frac{\mathbf{Z}_{k,k}+\frac{1}{2}\sqrt{\sum_{i=1}^{N}(\mathbf{Z}_{i,k}+\mathbf{Z}_{k,i})^2}}{2}, \tag{A.10}$$

then $p\geq p_{\text{L}}>0$, which contradicts the optimality condition (A.8). It can be concluded that whenever τ_k is large enough, at any KKT point no column will have size equal to τ_k. Furthermore, it can be easily checked that $\tau>\max_k\theta_k$ is a sufficient condition. The proof is complete.

A.3 Convergence Proof of the NS-SymNMF Algorithm

In this section, we prove Theorem 6. The analysis consists of a series of lemmas.

Lemma 8. *Consider using the update rules* (2.7) – (2.9) *to solve* (1.1). *Then we have*

$$\begin{aligned}\|\mathbf{\Lambda}^{(t+1)}-\mathbf{\Lambda}^{(t)}\|_F^2\leq{}&3N^2\tau^2\|\mathbf{X}^{(t+1)}-\mathbf{X}^{(t)}\|_F^2+3\|\mathbf{X}^{(t)}(\mathbf{Y}^{(t)})^T-\mathbf{Z}\|_F^2\|\mathbf{Y}^{(t+1)}-\mathbf{Y}^{(t)}\|_F^2\\&+3N\tau\|\mathbf{X}^{(t)}(\mathbf{Y}^{(t+1)}-\mathbf{Y}^{(t)})^T\|_F^2\end{aligned} \tag{A.11}$$

Proof: The optimality condition of the $\mathbf{X}$ subproblem (2.8) is given by

$$(\mathbf{X}^{(t+1)}(\mathbf{Y}^{(t+1)})^T - \mathbf{Z})\mathbf{Y}^{(t+1)} + \rho(\mathbf{X}^{(t+1)} - \mathbf{Y}^{(t+1)} + \mathbf{\Lambda}^{(t)}/\rho) = 0. \tag{A.12}$$

Substituting (2.9) into (A.12), we have

$$\mathbf{\Lambda}^{(t+1)} = -(\mathbf{X}^{(t+1)}(\mathbf{Y}^{(t+1)})^T - \mathbf{Z})\mathbf{Y}^{(t+1)}. \tag{A.13}$$

Subtracting the same equation in iteration t, we have the successive difference of the dual matrix (A.15),

$$\begin{aligned}
\mathbf{\Lambda}^{(t+1)} - \mathbf{\Lambda}^{(t)} =& -\left[\mathbf{X}^{(t+1)}(\mathbf{Y}^{(t+1)})^T\mathbf{Y}^{(t+1)} - \mathbf{X}^{(t)}(\mathbf{Y}^{(t)})^T\mathbf{Y}^{(t)} - \mathbf{Z}(\mathbf{Y}^{(t+1)} - \mathbf{Y}^{(t)})\right] \\
=& -\big[(\mathbf{X}^{(t+1)} - \mathbf{X}^{(t)})(\mathbf{Y}^{(t+1)})^T\mathbf{Y}^{(t+1)} + \mathbf{X}^{(t)}\big((\mathbf{Y}^{(t+1)})^T\mathbf{Y}^{(t+1)} - (\mathbf{Y}^{(t)})^T\mathbf{Y}^{(t)}\big) \\
&+ \mathbf{Z}(\mathbf{Y}^{(t+1)} - \mathbf{Y}^{(t)})\big] \\
=& \mathbf{Z}(\mathbf{Y}^{(t+1)} - \mathbf{Y}^{(t)}) - (\mathbf{X}^{(t+1)} - \mathbf{X}^{(t)})(\mathbf{Y}^{(t+1)})^T\mathbf{Y}^{(t+1)} \\
&\underbrace{- \frac{1}{2}\big(\mathbf{X}^{(t)}\big((\mathbf{Y}^{(t+1)} + \mathbf{Y}^{(t)})^T(\mathbf{Y}^{(t+1)} - \mathbf{Y}^{(t)}) + (\mathbf{Y}^{(t+1)} - \mathbf{Y}^{(t)})^T(\mathbf{Y}^{(t+1)} + \mathbf{Y}^{(t)})\big)\big)}_{\triangleq \mathcal{Q}}.
\end{aligned} \tag{A.14}$$

Note that the following is true

$$\begin{aligned}
\mathcal{Q} =& \frac{1}{2}\big(\mathbf{X}^{(t)}(\mathbf{Y}^{(t+1)} - \mathbf{Y}^{(t)})^T(\mathbf{Y}^{(t+1)} - \mathbf{Y}^{(t)}) + 2\mathbf{X}^{(t)}(\mathbf{Y}^{(t)})^T(\mathbf{Y}^{(t+1)} - \mathbf{Y}^{(t)})\big) \\
&+ \frac{1}{2}\mathbf{X}^{(t)}(\mathbf{Y}^{(t+1)} - \mathbf{Y}^{(t)})^T(\mathbf{Y}^{(t+1)} + \mathbf{Y}^{(t)}) \\
=& \mathbf{X}^{(t)}(\mathbf{Y}^{(t)})^T(\mathbf{Y}^{(t+1)} - \mathbf{Y}^{(t)}) + \mathbf{X}^{(t)}(\mathbf{Y}^{(t+1)} - \mathbf{Y}^{(t)})^T\mathbf{Y}^{(t+1)}.
\end{aligned} \tag{A.15}$$

Plugging (A.16) into (A.15), we have

$$\begin{aligned}
\mathbf{\Lambda}^{(t+1)} - \mathbf{\Lambda}^{(t)} =& \mathbf{Z}(\mathbf{Y}^{(t+1)} - \mathbf{Y}^{(t)}) - (\mathbf{X}^{(t+1)} - \mathbf{X}^{(t)})(\mathbf{Y}^{(t+1)})^T\mathbf{Y}^{(t+1)} \\
&- \mathbf{X}^{(t)}(\mathbf{Y}^{(t)})^T(\mathbf{Y}^{(t+1)} - \mathbf{Y}^{(t)}) - \mathbf{X}^{(t)}(\mathbf{Y}^{(t+1)} - \mathbf{Y}^{(t)})\mathbf{Y}^{(t+1)} \\
=& (\mathbf{Z} - \mathbf{X}^{(t)}(\mathbf{Y}^{(t)})^T)(\mathbf{Y}^{(t+1)} - \mathbf{Y}^{(t)}) - (\mathbf{X}^{(t+1)} - \mathbf{X}^{(t)})(\mathbf{Y}^{(t+1)})^T\mathbf{Y}^{(t+1)} \\
&- \mathbf{X}^{(t)}(\mathbf{Y}^{(t+1)} - \mathbf{Y}^{(t)})^T\mathbf{Y}^{(t+1)}.
\end{aligned} \tag{A.16}$$

Using triangle inequality, we arrive at

$$\begin{aligned}\|\mathbf{\Lambda}^{(t+1)}-\mathbf{\Lambda}^{(t)}\|_F &\le \|\mathbf{X}^{(t+1)}-\mathbf{X}^{(t)}\|_F\|(\mathbf{Y}^{(t+1)})^T\mathbf{Y}^{(t+1)}\|_F \\ &+ \|\mathbf{X}^{(t)}(\mathbf{Y}^{(t)})^T-\mathbf{Z}\|_F\|\mathbf{Y}^{(t+1)}-\mathbf{Y}^{(t)}\|_F + \|\mathbf{X}^{(t)}(\mathbf{Y}^{(t+1)}-\mathbf{Y}^{(t)})^T\|_F\|\mathbf{Y}^{(t+1)}\|_F. \quad \text{(A.17)}\end{aligned}$$

Since $\|\mathbf{Y}_i\|^2 \le \tau$, we know that $\|\mathbf{Y}\|_F \le \sqrt{N\tau}$. Squaring both sides of (A.17), we obtain

$$\begin{aligned}\|\mathbf{\Lambda}^{(t+1)}-\mathbf{\Lambda}^{(t)}\|_F^2 &\le 3N^2\tau^2\|\mathbf{X}^{(t+1)}-\mathbf{X}^{(t)}\|_F^2 \\ &+ 3\|\mathbf{X}^{(t)}(\mathbf{Y}^{(t)})^T-\mathbf{Z}\|_F^2\|\mathbf{Y}^{(t+1)}-\mathbf{Y}^{(t)}\|_F^2 + 3N\tau\|\mathbf{X}^{(t)}(\mathbf{Y}^{(t+1)}-\mathbf{Y}^{(t)})^T\|_F^2.\end{aligned}$$

The claim is proved.

In the second step, we bound the successive difference of the augmented Lagrangian.

Lemma 9. *Consider using the update rules* (2.7)–(2.9). *If*

$$\rho > 6N\tau \quad \text{and} \quad \beta^{(t)} > \frac{6}{\rho}\|\mathbf{X}^{(t)}(\mathbf{Y}^{(t)})^T-\mathbf{Z}\|_F^2 - \rho, \tag{A.18}$$

we have

$$\begin{aligned}&\mathcal{L}(\mathbf{X}^{(t+1)},\mathbf{Y}^{(t+1)},\mathbf{\Lambda}^{(t+1)}) - \mathcal{L}(\mathbf{X}^{(t)},\mathbf{Y}^{(t)},\mathbf{\Lambda}^{(t)}) \\ &\le -c_1\|\mathbf{X}^{(t+1)}-\mathbf{X}^{(t)}\|_F^2 - c_2\|\mathbf{X}^{(t)}(\mathbf{Y}^{(t+1)}-\mathbf{Y}^{(t)})^T\|_F^2 - c_3\|\mathbf{Y}^{(t+1)}-\mathbf{Y}^{(t)}\|_F^2\end{aligned} \tag{A.19}$$

where $c_1, c_2, c_3 > 0$ are some positive constants.

Proof: We have the following descent estimate

$$\begin{aligned}&\mathcal{L}(\mathbf{X}^{(t+1)},\mathbf{Y}^{(t+1)},\mathbf{\Lambda}^{(t+1)}) - \mathcal{L}(\mathbf{X}^{(t)},\mathbf{Y}^{(t)},\mathbf{\Lambda}^{(t)}) \\ &= \underbrace{\mathcal{L}(\mathbf{X}^{(t)},\mathbf{Y}^{(t+1)},\mathbf{\Lambda}^{(t)}) - \mathcal{L}(\mathbf{X}^{(t)},\mathbf{Y}^{(t)},\mathbf{\Lambda}^{(t)})}_{\triangleq\mathcal{A}} + \underbrace{\mathcal{L}(\mathbf{X}^{(t+1)},\mathbf{Y}^{(t+1)},\mathbf{\Lambda}^{(t)}) - \mathcal{L}(\mathbf{X}^{(t)},\mathbf{Y}^{(t+1)},\mathbf{\Lambda}^{(t)})}_{\triangleq\mathcal{B}}\end{aligned} \tag{A.20}$$

$$+ \underbrace{\mathcal{L}(\mathbf{X}^{(t+1)},\mathbf{Y}^{(t+1)},\mathbf{\Lambda}^{(t+1)}) - \mathcal{L}(\mathbf{X}^{(t+1)},\mathbf{Y}^{(t+1)},\mathbf{\Lambda}^{(t)})}_{\triangleq\mathcal{C}}$$

$$\le \underbrace{\widehat{\mathcal{L}}(\mathbf{X}^{(t)},\mathbf{Y}^{(t+1)},\mathbf{\Lambda}^{(t)}) - \mathcal{L}(\mathbf{X}^{(t)},\mathbf{Y}^{(t)},\mathbf{\Lambda}^{(t)})}_{\triangleq\widehat{\mathcal{A}}} + \mathcal{B} + \mathcal{C} \tag{A.21}$$

where

$$\widehat{\mathcal{L}}(\mathbf{X}^{(t)},\mathbf{Y},\mathbf{\Lambda}^{(t)})=\frac{1}{2}\|\mathbf{X}^{(t)}\mathbf{Y}^T-\mathbf{Z}\|_F^2+\frac{\rho}{2}\|\mathbf{X}^{(t)}-\mathbf{Y}+\mathbf{\Lambda}^{(t)}/\rho\|_F^2+\frac{\beta^{(t)}}{2}\|\mathbf{Y}-\mathbf{Y}^{(t)}\|_F^2, \tag{A.22}$$

which is an upper bound of $\mathcal{L}(\mathbf{X}^{(t)},\mathbf{Y},\mathbf{\Lambda}^{(t)})$. Next we bound different quantities in (A.21)

$$\begin{aligned}
\widehat{\mathcal{A}}=&\frac{1}{2}\|\mathbf{X}^{(t)}(\mathbf{Y}^{(t+1)})^T-\mathbf{Z}\|_F^2-\frac{1}{2}\|\mathbf{X}^{(t)}(\mathbf{Y}^{(t)})^T-\mathbf{Z}\|_F^2+\frac{\rho}{2}\|\mathbf{X}^{(t)}-\mathbf{Y}^{(t+1)}+\mathbf{\Lambda}^{(t)}/\rho\|_F^2\\
&-\frac{\rho}{2}\|\mathbf{X}^{(t)}-\mathbf{Y}^{(t)}+\mathbf{\Lambda}^{(t)}/\rho\|_F^2+\frac{\beta^{(t)}}{2}\|\mathbf{Y}^{(t+1)}-\mathbf{Y}^{(t)}\|_F^2\\
\overset{(a)}{=}&\langle(\mathbf{X}^{(t)}(\mathbf{Y}^{(t+1)})^T-\mathbf{Z})\mathbf{X}^{(t)},\mathbf{Y}^{(t+1)}-\mathbf{Y}^{(t)}\rangle-\frac{1}{2}\|\mathbf{X}^{(t)}(\mathbf{Y}^{(t+1)}-\mathbf{Y}^{(t)})^T\|_F^2\\
&+\rho\langle\mathbf{X}^{(t)}-\mathbf{Y}^{(t+1)}+\mathbf{\Lambda}^{(t)}/\rho,\mathbf{Y}^{(t+1)}-\mathbf{Y}^{(t)}\rangle-\frac{\rho}{2}\|\mathbf{Y}^{(t+1)}-\mathbf{Y}^{(t)}\|_F^2+\frac{\beta^{(t)}}{2}\|\mathbf{Y}^{(t+1)}-\mathbf{Y}^{(t)}\|_F^2\\
\overset{(b)}{\leq}&-\frac{1}{2}\|\mathbf{X}^{(t)}(\mathbf{Y}^{(t+1)}-\mathbf{Y}^{(t)})^T\|_F^2-\frac{\rho}{2}\|\mathbf{Y}^{(t+1)}-\mathbf{Y}^{(t)}\|_F^2-\frac{\beta^{(t)}}{2}\|\mathbf{Y}^{(t+1)}-\mathbf{Y}^{(t)}\|_F^2
\end{aligned}$$

where (a) due to the fact that Taylor expansion for quadratic problems is exact; (b) due to the optimality condition for problem (2.7).

Similarly, we have

$$\mathcal{B}\leq-\frac{1}{2}\|(\mathbf{X}^{(t+1)}-\mathbf{X}^{(t)})(\mathbf{Y}^{(t+1)})^T\|_F^2-\frac{\rho}{2}\|\mathbf{X}^{(t+1)}-\mathbf{X}^{(t)}\|_F^2, \tag{A.23}$$

$$\begin{aligned}
\mathcal{C}=&\langle\mathbf{X}^{(t+1)}-\mathbf{Y}^{(t+1)},\mathbf{\Lambda}^{(t+1)}-\mathbf{\Lambda}^{(t)}\rangle\\
\overset{(a)}{=}&\frac{1}{\rho}\|\mathbf{\Lambda}^{(t+1)}-\mathbf{\Lambda}^{(t)}\|_F^2
\end{aligned} \tag{A.24}$$

where (a) is from (2.9).

Substituting the result of Lemma 8 into (A.24), we can obtain

$$\begin{aligned}
&\mathcal{L}(\mathbf{X}^{(t+1)},\mathbf{Y}^{(t+1)},\mathbf{\Lambda}^{(t+1)})-\mathcal{L}(\mathbf{X}^{(t)},\mathbf{Y}^{(t)},\mathbf{\Lambda}^{(t)})\\
\leq&-\left(\frac{\rho}{2}-\frac{3N^2\tau^2}{\rho}\right)\|\mathbf{X}^{(t+1)}-\mathbf{X}^{(t)}\|_F^2-\left(\frac{1}{2}-\frac{3N\tau}{\rho}\right)\|\mathbf{X}^{(t)}(\mathbf{Y}^{(t+1)}-\mathbf{Y}^{(t)})^T\|_F^2\\
&-\left(\frac{\rho}{2}+\frac{\beta^{(t)}}{2}-\frac{3\|\mathbf{X}^{(t)}(\mathbf{Y}^{(t)})^T-\mathbf{Z}\|_F^2}{\rho}\right)\|\mathbf{Y}^{(t+1)}-\mathbf{Y}^{(t)}\|_F^2-\frac{1}{2}\|(\mathbf{X}^{(t+1)}-\mathbf{X}^{(t)})(\mathbf{Y}^{(t+1)})^T\|_F^2.
\end{aligned} \tag{A.25}$$

Therefore, from (A.25) if

$$\frac{\rho}{2} - \frac{3N^2\tau^2}{\rho} > 0, \tag{A.26a}$$

$$\frac{1}{2} - \frac{3N\tau}{\rho} > 0, \tag{A.26b}$$

$$\frac{\rho + \beta^{(t)}}{2} - \frac{3\|\mathbf{X}^{(t)}(\mathbf{Y}^{(t)})^T - \mathbf{Z}\|_F^2}{\rho} > 0, \tag{A.26c}$$

which are equivalent to

$$\rho > 6N\tau \quad \text{and} \quad \beta^{(t)} > \frac{6\|\mathbf{X}^{(t)}(\mathbf{Y}^{(t)})^T - \mathbf{Z}\|_F^2 - \rho^2}{\rho}, \tag{A.27}$$

we can have $\mathcal{L}(\mathbf{X}^{(t+1)}, \mathbf{Y}^{(t+1)}, \mathbf{\Lambda}^{(t+1)}) - \mathcal{L}(\mathbf{X}^{(t)}, \mathbf{Y}^{(t)}, \mathbf{\Lambda}^{(t)}) < 0$.

Then, it is concluded that $\mathcal{L}(\mathbf{X}^{(t+1)}, \mathbf{Y}^{(t+1)}, \mathbf{\Lambda}^{(t+1)})$ is decreasing.

In the next step we prove that $\mathcal{L}(\mathbf{X}^{(t+1)}, \mathbf{Y}^{(t+1)}, \mathbf{\Lambda}^{(t+1)})$ is lower bounded.

Lemma 10. *Consider using the update rules (2.7) (2.8) (2.9). If $\rho \geq N\tau$ is satisfied, we have*

$$\mathcal{L}(\mathbf{X}^{(t+1)}, \mathbf{Y}^{(t+1)}, \mathbf{\Lambda}^{(t+1)}) \geq 0. \tag{A.28}$$

Proof: At iteration $t+1$, the augmented Lagrangian can be lower bounded as

$$\begin{aligned}
&\mathcal{L}(\mathbf{X}^{(t+1)}, \mathbf{Y}^{(t+1)}, \mathbf{\Lambda}^{(t+1)}) \\
=&\frac{1}{2}\|\mathbf{X}^{(t+1)}(\mathbf{Y}^{(t+1)})^T - \mathbf{Z}\|_F^2 + \langle \mathbf{X}^{(t+1)} - \mathbf{Y}^{(t+1)}, \mathbf{\Lambda}^{(t+1)} \rangle + \frac{\rho}{2}\|\mathbf{X}^{(t+1)} - \mathbf{Y}^{(t+1)}\|_F^2 \\
\overset{(a)}{=}&\frac{1}{2}\|\mathbf{X}^{(t+1)}(\mathbf{Y}^{(t+1)})^T - \mathbf{Z}\|_F^2 + \langle \mathbf{X}^{(t+1)} - \mathbf{Y}^{(t+1)}, -(\mathbf{X}^{(t+1)}(\mathbf{Y}^{(t+1)})^T - \mathbf{Z})\mathbf{Y}^{(t+1)} \rangle \\
&+ \frac{\rho}{2}\|\mathbf{X}^{(t+1)} - \mathbf{Y}^{(t+1)}\|_F^2 \\
\overset{(b)}{\geq}&\frac{1}{2}(\rho - N\tau)\|\mathbf{X}^{(t+1)} - \mathbf{Y}^{(t+1)}\|_F^2
\end{aligned} \tag{A.29}$$

where (a) due to (A.13); (b) because the fact that

$$\begin{aligned}
0 \leq&\|(\mathbf{X}^{(t+1)} - \mathbf{Y}^{(t+1)})(\mathbf{Y}^{(t+1)})^T - (\mathbf{X}^{(t+1)}(\mathbf{Y}^{(t+1)})^T - \mathbf{Z})\|_F^2 \\
=&\|(\mathbf{X}^{(t+1)} - \mathbf{Y}^{(t+1)})(\mathbf{Y}^{(t+1)})^T\|_F^2 - 2\langle (\mathbf{Y}^{(t+1)})^T(\mathbf{X}^{(t+1)} - \mathbf{Y}^{(t+1)}), \mathbf{X}^{(t+1)}(\mathbf{Y}^{(t+1)})^T - \mathbf{Z} \rangle \\
&+ \|\mathbf{X}^{(t+1)}(\mathbf{Y}^{(t+1)})^T - \mathbf{Z})\|_F^2,
\end{aligned}$$

and $\|\mathbf{Y}\|_F^2 \leq N\tau$.

From (A.29), we know that if $\rho \geq N\tau$, we have $\mathcal{L}(\mathbf{X}^{(t+1)}, \mathbf{Y}^{(t+1)}, \mathbf{\Lambda}^{(t+1)}) \geq 0$.

These lemmas lead to the main convergence claim.

Proof: Combing (A.19) and (A.28), we have

$$\begin{aligned} &\lim_{t\to\infty} \|\mathbf{X}^{(t+1)} - \mathbf{X}^{(t)}\|_F = 0, \\ &\lim_{t\to\infty} \|\mathbf{X}^{(t)}(\mathbf{Y}^{(t+1)} - \mathbf{Y}^{(t)})^T\| = 0, \\ &\lim_{t\to\infty} \|\mathbf{X}^{(t)}(\mathbf{Y}^{(t)})^T - \mathbf{Z}\|_F^2 \|\mathbf{Y}^{(t+1)} - \mathbf{Y}^{(t)}\|_F = 0. \end{aligned} \tag{A.30}$$

By Lemma 8, we have

$$\lim_{t\to\infty} \|\mathbf{\Lambda}^{(t+1)} - \mathbf{\Lambda}^{(t)}\|_F = 0, \tag{A.31}$$

which implies $\lim_{t\to\infty} \|\mathbf{X}^{(t)} - \mathbf{Y}^{(t)}\|_F = 0$. Combining with (A.30), we can further know that $\lim_{t\to\infty} \|\mathbf{Y}^{(t+1)} - \mathbf{Y}^{(t)}\|_F = 0$. The boundedness assumption of $\mathbf{X}^{(t)}$ then follows from the boundedness of $\mathbf{Y}^{(t)}$. Using the expression of $\mathbf{\Lambda}^{(t)}$ in (A.13), one can show that $\{\mathbf{\Lambda}^{(t)}\}$ is also bounded.

The optimality condition of (2.7) is given by

$$\begin{aligned} \big\langle (\mathbf{X}^{(t)})^T(\mathbf{X}^{(t)}(\mathbf{Y}^{(t+1)})^T - \mathbf{Z}) - \rho(\mathbf{X}^{(t)} - \mathbf{Y}^{(t+1)} + \mathbf{\Lambda}^{(t)}/\rho)^T \\ + \beta^{(t)}(\mathbf{Y}^{(t+1)} - \mathbf{Y}^{(t)})^T, (\mathbf{Y} - \mathbf{Y}^{(t+1)})^T \big\rangle \geq 0, \qquad \forall\, \mathbf{Y} \geq 0 \quad \text{and} \quad \|\mathbf{Y}_i\|^2 \leq \tau\ \forall i. \end{aligned} \tag{A.32}$$

Substituting (A.13) into (A.32), using (A.30), and taking limit over any converging subsequence of $(\mathbf{X}^{(t)}, \mathbf{Y}^{(t)}; \mathbf{\Lambda}^{(t)})$, we have

$$\begin{aligned} \langle (\mathbf{X}^*)^T(\mathbf{X}^*(\mathbf{Y}^*)^T - \mathbf{Z}) + ((\mathbf{X}^*(\mathbf{Y}^*)^T - \mathbf{Z})\mathbf{Y}^*)^T - \rho(\mathbf{X}^* - \mathbf{Y}^*)^T, (\mathbf{Y} - \mathbf{Y}^*)^T \rangle \geq 0, \\ \forall\, \mathbf{Y} \geq 0 \quad \text{and} \quad \|\mathbf{Y}_i\|^2 \leq \tau\ \forall i. \end{aligned} \tag{A.33}$$

The optimality condition of (2.8) is given by

$$(\mathbf{X}^{(t+1)}(\mathbf{Y}^{(t+1)})^T - \mathbf{Z})(\mathbf{Y}^{(t+1)}) + \rho(\mathbf{X}^{(t+1)} - \mathbf{Y}^{(t+1)} + \mathbf{\Lambda}^{(t)}/\rho) = 0. \tag{A.34}$$

Taking limit of (A.34) over the same subsequence, we have

$$(\mathbf{X}^*(\mathbf{Y}^*)^T - \mathbf{Z})\mathbf{Y}^* + \rho(\mathbf{X}^* - \mathbf{Y}^* + \mathbf{\Lambda}^*/\rho) = 0. \tag{A.35}$$

Using the fact $\mathbf{X}^* = \mathbf{Y}^*$, we have

$$\begin{aligned}
&\left\langle (\mathbf{X}^*(\mathbf{X}^*)^T - \frac{\mathbf{Z}^T + \mathbf{Z}}{2})\mathbf{X}^*, \mathbf{X} - \mathbf{X}^* \right\rangle \geq 0, \qquad \forall\, \mathbf{X} \geq 0,\ \|\mathbf{X}_i\|^2 \leq \tau\ \forall i, \\
&(\mathbf{X}^*(\mathbf{X}^*)^T - \mathbf{Z})\mathbf{X}^* + \mathbf{\Lambda}^* = 0,
\end{aligned} \tag{A.36}$$

which are the KKT conditions of problem (1.1).

A.4 Convergence Rate Proof of the NS-SymNMF Algorithm

Proof: First, from Theorem 6 we know that $\|\mathbf{X}^{(t)}\|_F$ is bounded, then there must exist a finite $\gamma > 0$ such that $\|\mathbf{X}^{(t)}\|_F^2 \leq N\gamma, \forall t$, where γ is only dependent on τ, N and $\|\mathbf{Z}\|_F$.

From the optimality condition of $\mathbf{Y}$ in (2.7), we have

$$\begin{aligned}
(\mathbf{Y}^{(t+1)})^T = \mathsf{proj}_{\mathcal{Y}}\Big[&(\mathbf{Y}^{(t+1)})^T - ((\mathbf{X}^{(t)})^T(\mathbf{X}^{(t)}(\mathbf{Y}^{(t+1)})^T - \mathbf{Z}) - \rho(\mathbf{X}^{(t)} \\
&- \mathbf{Y}^{(t+1)} + \mathbf{\Lambda}^{(t)}/\rho)^T + \beta^{(t)}(\mathbf{Y}^{(t+1)} - \mathbf{Y}^{(t)})^T)\Big].
\end{aligned}$$

Then, we have

$$\begin{aligned}
&\left\|(\mathbf{Y}^{(t)})^T - \mathsf{proj}_{\mathcal{Y}}\big[(\mathbf{Y}^{(t)})^T - ((\mathbf{X}^{(t)})^T(\mathbf{X}^{(t)}(\mathbf{Y}^{(t)})^T - \mathbf{Z}) - \rho(\mathbf{X}^{(t)} - \mathbf{Y}^{(t)} + \mathbf{\Lambda}^{(t)}/\rho)^T)\big]\right\|_F \\
&= \Big\|(\mathbf{Y}^{(t)})^T - (\mathbf{Y}^{(t+1)})^T + (\mathbf{Y}^{(t+1)})^T \\
&\quad - \mathsf{proj}_{\mathcal{Y}}\big[(\mathbf{Y}^{(t)})^T - ((\mathbf{X}^{(t)})^T(\mathbf{X}^{(t)}(\mathbf{Y}^{(t)})^T - \mathbf{Z}) \quad - \rho(\mathbf{X}^{(t)} - \mathbf{Y}^{(t)} + \mathbf{\Lambda}^{(t)}/\rho)^T)\big]\Big\|_F \\
&\overset{(a)}{\leq} \|\mathbf{Y}^{(t)} - \mathbf{Y}^{(t+1)}\|_F \\
&\quad + \Big\|\mathsf{proj}_{\mathcal{Y}}\big[(\mathbf{Y}^{(t+1)})^T - ((\mathbf{X}^{(t)})^T(\mathbf{X}^{(t)}(\mathbf{Y}^{(t+1)})^T - \mathbf{Z}) \\
&\quad - \rho(\mathbf{X}^{(t)} - \mathbf{Y}^{(t+1)} + \mathbf{\Lambda}^{(t)}/\rho)^T + \beta^{(t)}(\mathbf{Y}^{(t+1)} - \mathbf{Y}^{(t)})^T)\big] \\
&\quad - \mathsf{proj}_{\mathcal{Y}}\big[(\mathbf{Y}^{(t)})^T - ((\mathbf{X}^{(t)})^T(\mathbf{X}^{(t)}(\mathbf{Y}^{(t)})^T - \mathbf{Z}) - \rho(\mathbf{X}^{(t)} - \mathbf{Y}^{(t)} + \mathbf{\Lambda}^{(t)}/\rho)^T)\big]\Big\|_F \\
&\overset{(b)}{\leq} (2 + \rho + \beta^{(t)})\|\mathbf{Y}^{(t+1)} - \mathbf{Y}^{(t)}\|_F + \|(\mathbf{X}^{(t)})^T\mathbf{X}^{(t)}(\mathbf{Y}^{(t+1)} - \mathbf{Y}^{(t)})^T\|_F \\
&\overset{(c)}{\leq} (2 + \rho + \beta^{(t)})\|\mathbf{Y}^{(t+1)} - \mathbf{Y}^{(t)}\|_F + \sqrt{N\gamma}\|\mathbf{X}^{(t)}(\mathbf{Y}^{(t+1)} - \mathbf{Y}^{(t)})^T\|_F
\end{aligned} \tag{A.37}$$

where $\mathsf{proj}_{\mathcal{Y}}$ denotes the projection of $\mathbf{Y}$ to the feasible space; in (a) we used triangle inequality; (b) is due to the nonexpansiveness of the projection operator; and (c) is because of the boundedness of $\|\mathbf{X}\|_F$.

Similarly, we can bound the size of the gradient of the augmented Lagrangian with respect to $\mathbf{X}$ by the following series of inequalities

$$\begin{aligned}
&\|\nabla_{\mathbf{X}}\mathcal{L}(\mathbf{X}^{(t)},\mathbf{Y}^{(t)},\mathbf{\Lambda}^{(t)})\|_F = \|(\mathbf{X}^{(t)}(\mathbf{Y}^{(t)})^T-\mathbf{Z})\mathbf{Y}^{(t)}+\rho(\mathbf{X}^{(t)}-\mathbf{Y}^{(t)}+\mathbf{\Lambda}^{(t)}/\rho)\|_F\\
&\overset{(a)}{=} \big\|(\mathbf{X}^{(t)}(\mathbf{Y}^{(t)})^T-\mathbf{Z})\mathbf{Y}^{(t)}+\rho(\mathbf{X}^{(t)}-\mathbf{Y}^{(t)}+\mathbf{\Lambda}^{(t)}/\rho)\\
&\quad -((\mathbf{X}^{(t+1)}(\mathbf{Y}^{(t+1)})^T-\mathbf{Z})\mathbf{Y}^{(t+1)} \quad +\rho(\mathbf{X}^{(t+1)}-\mathbf{Y}^{(t+1)}+\mathbf{\Lambda}^{(t)}/\rho))\big\|_F\\
&\le \|(\mathbf{X}^{(t)}(\mathbf{Y}^{(t)})^T-\mathbf{Z})\mathbf{Y}^{(t)}-((\mathbf{X}^{(t+1)}(\mathbf{Y}^{(t+1)})^T-\mathbf{Z})\mathbf{Y}^{(t+1)})\|_F\\
&\quad +\rho\|\mathbf{Y}^{(t+1)}-\mathbf{Y}^{(t)}\|_F+\rho\|\mathbf{X}^{(t+1)}-\mathbf{X}^{(t)}\|_F\\
&\overset{(b)}{=} \|\mathbf{\Lambda}^{(t+1)}-\mathbf{\Lambda}^{(t)}\|_F+\rho\|\mathbf{Y}^{(t+1)}-\mathbf{Y}^{(t)}\|_F+\rho\|\mathbf{X}^{(t+1)}-\mathbf{X}^{(t)}\|_F
\end{aligned}\tag{A.38}$$

where (a) is from the optimality condition of the $\mathbf{X}$ subproblem (A.12); (b) is true due to (A.14) and (A.13). Squaring both sides of (A.38) and applying Lemma 8, we have

$$\begin{aligned}
\|\nabla_{\mathbf{X}}\mathcal{L}(\mathbf{X}^{(t)},\mathbf{Y}^{(t)},\mathbf{\Lambda}^{(t)})\|_F^2 &\le 3(3N^2\tau^2+\rho^2)\|\mathbf{X}^{(t+1)}-\mathbf{X}^{(t)}\|_F^2\\
&+3(3\|\mathbf{X}^{(t)}(\mathbf{Y}^{(t)})^T-\mathbf{Z}\|_F^2+\rho^2)\|\mathbf{Y}^{(t+1)}-\mathbf{Y}^{(t)}\|_F^2\\
&+9N\tau\|\mathbf{X}^{(t)}(\mathbf{Y}^{(t+1)}-\mathbf{Y}^{(t)})^T\|_F^2.
\end{aligned}\tag{A.39}$$

Due to the boundedness of $\mathbf{X}^{(t)}$ and $\mathbf{Y}^{(t)}$, we must have that for some $\delta>0$, $\|\mathbf{X}^{(t)}(\mathbf{Y}^{(t)})^T-\mathbf{Z}\|_F\le\delta$.

Therefore, combining (A.37) and (A.39), there must exists a finite positive number σ_1 such that

$$\|\widetilde{\nabla}\mathcal{L}(\mathbf{X}^{(t)},\mathbf{Y}^{(t)},\mathbf{\Lambda}^{(t)})\|_F^2\le\sigma_1\mathcal{F}\tag{A.40}$$

where

$$\mathcal{F}\triangleq\|\mathbf{X}^{(t+1)}-\mathbf{X}^{(t)}\|_F^2+\|\mathbf{Y}^{(t+1)}-\mathbf{Y}^{(t)}\|_F^2+\|\mathbf{X}^{(t)}(\mathbf{Y}^{(t+1)}-\mathbf{Y}^{(t)})^T\|_F^2\tag{A.41}$$

In particular, we have $\sigma_1\triangleq\max\{3(3N^2\tau^2+\rho^2),3(2+\rho+\beta^{(t)})^2+3(3\delta^2+\rho^2),3\gamma+9N\tau\}$ and $\beta^{(t)}\le 6\delta^2/\rho$.

According to Lemma 8, we have

$$\|\mathbf{X}^{(t+1)} - \mathbf{Y}^{(t+1)}\|_F^2 = \frac{1}{\rho^2}\|\mathbf{\Lambda}^{(t+1)} - \mathbf{\Lambda}^{(t)}\|_F^2 \leq \sigma_2 \mathcal{F} \tag{A.42}$$

where some constant $\sigma_2 \triangleq \max\{3N^2\tau^2/\rho^2, 3\delta^2/\rho^2, 3N\tau/\rho^2\}$.

Also, we have

$$\begin{aligned}\|\mathbf{X}^{(t)} - \mathbf{Y}^{(t)}\|_F =& \|\mathbf{X}^{(t)} - \mathbf{X}^{(t+1)} + \mathbf{X}^{(t+1)} - \mathbf{Y}^{(t+1)} + \mathbf{Y}^{(t+1)} - \mathbf{Y}^{(t)}\|_F \\ \leq& \|\mathbf{X}^{(t)} - \mathbf{X}^{(t+1)}\|_F + \|\mathbf{X}^{(t+1)} - \mathbf{Y}^{(t+1)}\|_F + \|\mathbf{Y}^{(t+1)} - \mathbf{Y}^{(t)}\|_F,\end{aligned} \tag{A.43}$$

which yields

$$\|\mathbf{X}^{(t)} - \mathbf{Y}^{(t)}\|_F^2 \leq \sigma_3 \mathcal{F} \tag{A.44}$$

for some constant $\sigma_3 \triangleq \max\{9N^2\tau^2/\rho^2 + 3, 9\delta^2/\rho^2 + 3, 9N\tau/\rho^2\}$.

The inequalities (A.40) and (A.44) imply that

$$\|\widetilde{\nabla}\mathcal{L}(\mathbf{X}^{(t)}, \mathbf{Y}^{(t)}, \mathbf{\Lambda}^{(t)})\|_F^2 + \|\mathbf{X}^{(t)} - \mathbf{Y}^{(t)}\|_F^2 \leq (\sigma_1 + \sigma_3)\mathcal{F}. \tag{A.45}$$

According to Lemma 9, there exists a constant $\sigma_4 \triangleq \min\{c_1, c_2, c_3\}$ such that

$$\mathcal{L}(\mathbf{X}^{(t)}, \mathbf{Y}^{(t)}, \mathbf{\Lambda}^{(t)}) - \mathcal{L}(\mathbf{X}^{(t+1)}, \mathbf{Y}^{(t+1)}, \mathbf{\Lambda}^{(t+1)}) \geq \sigma_4 \mathcal{F}. \tag{A.46}$$

Combining (A.45) and (A.46), we have

$$\begin{aligned}&\|\widetilde{\nabla}\mathcal{L}(\mathbf{X}^{(t)}, \mathbf{Y}^{(t)}, \mathbf{\Lambda}^{(t)}\|_F^2 + \|\mathbf{X}^{(t)} - \mathbf{Y}^{(t)}\|_F^2 \leq \\ &\qquad \frac{\sigma_1 + \sigma_3}{\sigma_4}(\mathcal{L}(\mathbf{X}^{(t)}, \mathbf{Y}^{(t)}, \mathbf{\Lambda}^{(t)}) - \mathcal{L}(\mathbf{X}^{(t+1)}, \mathbf{Y}^{(t+1)}, \mathbf{\Lambda}^{(t+1)})).\end{aligned} \tag{A.47}$$

Summing both sides of (A.47) over $t = 1, \ldots, r$, we have

$$\begin{aligned}&\sum_{t=1}^{r} \|\widetilde{\nabla}\mathcal{L}(\mathbf{X}^{(t)}, \mathbf{Y}^{(t)}, \mathbf{\Lambda}^{(t)})\|_F^2 + \|\mathbf{X}^{(t)} - \mathbf{Y}^{(t)}\|_F^2 \\ \leq& \frac{\sigma_1 + \sigma_3}{\sigma_4}(\mathcal{L}(\mathbf{X}^{(1)}, \mathbf{Y}^{(1)}, \mathbf{\Lambda}^{(1)}) - \mathcal{L}(\mathbf{X}^{(t+1)}, \mathbf{Y}^{(t+1)}, \mathbf{\Lambda}^{(t+1)})) \\ \overset{(a)}{\leq}& \frac{\sigma_1 + \sigma_3}{\sigma_4}\mathcal{L}(\mathbf{X}^{(1)}, \mathbf{Y}^{(1)}, \mathbf{\Lambda}^{(1)})\end{aligned} \tag{A.48}$$

where (a) due to Lemma 10.

According to the definition of $T(\epsilon)$ and $\mathcal{P}(\mathbf{X}^{(t)}, \mathbf{Y}^{(t)}, \mathbf{\Lambda}^{(t)})$, the above inequality becomes

$$T(\epsilon)\epsilon \leq \frac{\sigma_1 + \sigma_3}{\sigma_4}\mathcal{L}(\mathbf{X}^{(1)}, \mathbf{Y}^{(1)}, \mathbf{\Lambda}^{(1)}). \tag{A.49}$$

Dividing both sides by $T(\epsilon)$, and by setting $C \triangleq (\sigma_1 + \sigma_3)/\sigma_4$, the desired result is obtained.

A.5 Sufficient Condition of Optimality of SymNMF

Proof: Let $\mathbf{\Omega}$ be the matrix of Lagrange multipliers. The Lagrangian of (1.1) is given by

$$\mathcal{L}(\mathbf{X}, \mathbf{\Omega}) = \frac{1}{2}\mathrm{Tr}\left((\mathbf{X}\mathbf{X}^T - \mathbf{Z})^T(\mathbf{X}\mathbf{X}^T - \mathbf{Z})\right) - \langle \mathbf{X}, \mathbf{\Omega}\rangle. \tag{A.50}$$

Let $(\mathbf{X}^*, \mathbf{\Omega}^*)$ be a KKT point of (1.1). To show global optimality of $(\mathbf{X}^*, \mathbf{\Omega}^*)$, it is sufficient to prove the following saddle point condition (72, pp. 238)

$$\mathcal{L}(\mathbf{X}^*, \mathbf{\Omega}) \leq \mathcal{L}(\mathbf{X}^*, \mathbf{\Omega}^*) \leq \mathcal{L}(\mathbf{X}, \mathbf{\Omega}^*),\ \forall\ \mathbf{\Omega} \geq 0,\ \forall\ \mathbf{X}. \tag{A.51}$$

To show the left hand side of (A.51), we have the following

$$\mathcal{L}(\mathbf{X}^*, \mathbf{\Omega}^*) - \mathcal{L}(\mathbf{X}^*, \mathbf{\Omega}) = -\langle \mathbf{X}^*, \mathbf{\Omega}^*\rangle - (-\langle \mathbf{X}^*, \mathbf{\Omega}\rangle) = \langle \mathbf{X}^*, \mathbf{\Omega} - \mathbf{\Omega}^*\rangle \overset{(a)}{=} \langle \mathbf{X}^*, \mathbf{\Omega}\rangle \overset{(b)}{\geq} 0. \tag{A.52}$$

where (a) due to (2.4d); (b) due to $\mathbf{\Omega} \geq 0$ and (2.4c).

Next we show the right hand side of (A.51)

$$\begin{aligned}
\mathcal{L}(\mathbf{X}, \mathbf{\Omega}^*) - \mathcal{L}(\mathbf{X}^*, \mathbf{\Omega}^*) =& \underbrace{\frac{1}{2}\mathrm{Tr}[(\mathbf{X}\mathbf{X}^T - \mathbf{X}^*(\mathbf{X}^*)^T)(\mathbf{X}\mathbf{X}^T - \mathbf{X}^*(\mathbf{X}^*)^T)]}_{\triangleq \mathcal{M}} \\
&+ \mathrm{Tr}[(\mathbf{X}^*(\mathbf{X}^*)^T - \mathbf{Z}^T)(\mathbf{X}\mathbf{X}^T - \mathbf{X}^*(\mathbf{X}^*)^T)] \\
&- \langle \mathbf{X} - \mathbf{X}^*, \mathbf{\Omega}^*\rangle \\
\overset{(a)}{\geq}& \langle \mathbf{X} - \mathbf{X}^*, \left(\mathbf{X}^*(\mathbf{X}^*)^T - \frac{\mathbf{Z}^T + \mathbf{Z}}{2}\right)(\mathbf{X} + \mathbf{X}^*)\rangle - \langle \mathbf{X} - \mathbf{X}^*, \mathbf{\Omega}^*\rangle \qquad \text{(A.53)} \\
\overset{(b)}{=}& \langle \mathbf{X} - \mathbf{X}^*, \left(\mathbf{X}^*(\mathbf{X}^*)^T - \frac{\mathbf{Z}^T + \mathbf{Z}}{2}\right)(\mathbf{X} - \mathbf{X}^*)\rangle \\
=& \mathrm{Tr}\big[(\mathbf{X} - \mathbf{X}^*)^T \underbrace{\left(\mathbf{X}^*(\mathbf{X}^*)^T - \frac{\mathbf{Z}^T + \mathbf{Z}}{2}\right)}_{\triangleq \mathbf{S}}(\mathbf{X} - \mathbf{X}^*)\big] \qquad \text{(A.54)}
\end{aligned}$$

where (a) due to $\mathcal{M} \geq 0$ and the fact that

$$\mathbf{X}\mathbf{X}^T - \mathbf{X}^*(\mathbf{X}^*)^T = \frac{1}{2}\big[(\mathbf{X}+\mathbf{X}^*)(\mathbf{X}-\mathbf{X}^*)^T + (\mathbf{X}-\mathbf{X}^*)(\mathbf{X}+\mathbf{X}^*)^T\big]; \tag{A.55}$$

(b) is true because of (2.4a). Clearly, if we have $\mathbf{S} \succeq 0$, then the following inequality must be true

$$\mathcal{L}(\mathbf{X}, \boldsymbol{\Omega}^*) - \mathcal{L}(\mathbf{X}^*, \boldsymbol{\Omega}^*) \geq 0.$$

This completes the proof.

A.6 Sufficient Local Optimality Condition

Proof: We first simplify the term $\mathcal{M}$ in (A.53) as follows.

$$\begin{aligned}
&\frac{1}{2}\mathrm{Tr}[(\mathbf{X}\mathbf{X}^T - \mathbf{X}^*(\mathbf{X}^*)^T)^T(\mathbf{X}\mathbf{X}^T - \mathbf{X}^*(\mathbf{X}^*)^T)] \\
\overset{(a)}{=}&\frac{1}{2}\mathrm{Tr}\Big[\left((\mathbf{X}-\mathbf{X}^*)\mathbf{X}^T + \mathbf{X}^*(\mathbf{X}-\mathbf{X}^*)^T\right)^T\left((\mathbf{X}-\mathbf{X}^*)\mathbf{X}^T + \mathbf{X}^*(\mathbf{X}-\mathbf{X}^*)^T\right)\Big] \\
\overset{(b)}{=}&\frac{1}{2}\mathrm{Tr}\Big[\left(\widehat{\mathbf{Y}}(\widehat{\mathbf{Y}}+\mathbf{X}^*)^T + \mathbf{X}^*\widehat{\mathbf{Y}}^T\right)^T\left(\widehat{\mathbf{Y}}(\widehat{\mathbf{Y}}+\mathbf{X}^*)^T + \mathbf{X}^*\widehat{\mathbf{Y}}^T\right)\Big] \\
\overset{(c)}{=}&\frac{1}{2}\mathrm{Tr}\big[\mathbf{U}^T\mathbf{U} + \mathbf{X}^*\widehat{\mathbf{Y}}^T\mathbf{U} + \widehat{\mathbf{Y}}(\mathbf{X}^*)^T\mathbf{U} + \mathbf{X}^*\widehat{\mathbf{Y}}^T\mathbf{U} \\
&+ \mathbf{X}^*\widehat{\mathbf{Y}}^T\widehat{\mathbf{Y}}(\mathbf{X}^*)^T + \mathbf{X}^*\widehat{\mathbf{Y}}^T\mathbf{X}^*\widehat{\mathbf{Y}}^T \\
&+ \widehat{\mathbf{Y}}(\mathbf{X}^*)^T\mathbf{U} + \widehat{\mathbf{Y}}(\mathbf{X}^*)^T\widehat{\mathbf{Y}}(\mathbf{X}^*)^T + \widehat{\mathbf{Y}}(\mathbf{X}^*)^T\mathbf{X}^*\widehat{\mathbf{Y}}^T\big] \\
=&\frac{1}{2}\mathrm{Tr}\left[\mathbf{U}\mathbf{U}^T + 4\mathbf{U}\mathbf{X}^*\widehat{\mathbf{Y}}^T + 2\widehat{\mathbf{Y}}(\mathbf{X}^*)^T\mathbf{X}^*\widehat{\mathbf{Y}}^T\right] + \mathrm{Tr}\left[\mathbf{X}^*\widehat{\mathbf{Y}}^T\mathbf{X}^*\widehat{\mathbf{Y}}^T\right] \\
=&\frac{1}{2}\mathrm{Tr}\left[\widehat{\mathbf{Y}}\begin{bmatrix}\widehat{\mathbf{Y}}^T & \mathbf{I}\end{bmatrix}\begin{bmatrix}\mathbf{I} & 4\mathbf{X}^* \\ \mathbf{0} & 2(\mathbf{X}^*)^T\mathbf{X}^*\end{bmatrix}\begin{bmatrix}\widehat{\mathbf{Y}}^T & \mathbf{I}\end{bmatrix}^T\widehat{\mathbf{Y}}^T\right] + \mathrm{Tr}\left[\mathbf{X}^*\widehat{\mathbf{Y}}^T\mathbf{X}^*\widehat{\mathbf{Y}}^T\right]
\end{aligned}$$

where (a) is due to the fact that

$$\mathbf{X}\mathbf{X}^T - \mathbf{X}^*(\mathbf{X}^*)^T = (\mathbf{X}-\mathbf{X}^*)\mathbf{X}^T + \mathbf{X}^*(\mathbf{X}-\mathbf{X}^*)^T; \tag{A.56}$$

in (b) we defined $\widehat{\mathbf{Y}} \triangleq \mathbf{X} - \mathbf{X}^*$ which shows the difference between $\mathbf{X}$ and $\mathbf{X}^*$; and in (c) we defined $\mathbf{U} \triangleq \widehat{\mathbf{Y}}\widehat{\mathbf{Y}}^T = \mathbf{U}^T$.

Combining (A.54) and (A.56), we have

$$\begin{aligned}\mathcal{L}(\mathbf{X},\mathbf{\Omega}^*)-\mathcal{L}(\mathbf{X}^*,\mathbf{\Omega}^*) =&\mathrm{Tr}\left[\widehat{\mathbf{Y}}\left[\frac{1}{2}\widehat{\mathbf{Y}}^T\widehat{\mathbf{Y}}+2\widehat{\mathbf{Y}}^T\mathbf{X}^*+(\mathbf{X}^*)^T\mathbf{X}^*\right]\widehat{\mathbf{Y}}^T\right]\\ &+\mathrm{Tr}\left[\mathbf{X}^*\widehat{\mathbf{Y}}^T\mathbf{X}^*\widehat{\mathbf{Y}}^T\right]+\mathrm{Tr}\left[\widehat{\mathbf{Y}}^T\left(\mathbf{X}^*(\mathbf{X}^*)^T-\frac{\mathbf{Z}^T+\mathbf{Z}}{2}\right)\widehat{\mathbf{Y}}\right]\\ =&\sum_m^K\sum_n^K(\widehat{\mathbf{Y}}'_m)^T\mathcal{K}_{m,n}\widehat{\mathbf{Y}}'_n+\sum_m^K\sum_n^K(\widehat{\mathbf{Y}}'_m)^T\widetilde{\mathcal{K}}_{m,n}\widehat{\mathbf{Y}}'_n+\sum_m^K(\widehat{\mathbf{Y}}'_m)^T\mathbf{S}\widehat{\mathbf{Y}}'_m\\ =&\mathrm{vec}(\widehat{\mathbf{Y}})^T\mathcal{T}\mathrm{vec}(\widehat{\mathbf{Y}})\end{aligned} \tag{A.57}$$

where

$$\mathcal{T}\triangleq\begin{bmatrix}\mathcal{K}_{1,1}\mathbf{I}+\widetilde{\mathcal{K}}_{1,1}+\mathbf{S} & \cdots & \mathcal{K}_{1,K}\mathbf{I}+\widetilde{\mathcal{K}}_{1,K}\\ \vdots & \cdots & \vdots\\ \mathcal{K}_{K,1}\mathbf{I}+\widetilde{\mathcal{K}}_{K,1} & \cdots & \mathcal{K}_{K,K}\mathbf{I}+\widetilde{\mathcal{K}}_{K,K}+\mathbf{S}\end{bmatrix}, \tag{A.58}$$

$$\mathcal{K}_{m,n}\triangleq\frac{1}{2}(\widehat{\mathbf{Y}}'_m)^T\widehat{\mathbf{Y}}'_n+2(\widehat{\mathbf{Y}}'_m)^T\mathbf{X}'^*_n+(\mathbf{X}'^*_m)^T\mathbf{X}'^*_n, \tag{A.59}$$

and $\widetilde{\mathcal{K}}_{m,n}\triangleq\mathbf{X}'^*_n(\mathbf{X}'^*_m)^T$, (m,n) denotes the (m,n)th block of a matrix, $\mathbf{X}'^*_m$ ($\widehat{\mathbf{Y}}'_n$) denotes the mth (or nth) *column* of the matrix $\mathbf{X}^*$ (or $\widehat{\mathbf{Y}}$).

For the (m,n)th block, we have

$$\begin{aligned}&(\widehat{\mathbf{Y}}'_m)^T\Big(\left(\frac{1}{2}(\widehat{\mathbf{Y}}'_m)^T\widehat{\mathbf{Y}}'_n+2(\widehat{\mathbf{Y}}'_m)^T\mathbf{X}'^*_n+(\mathbf{X}'^*_m)^T\mathbf{X}'^*_n\right)\mathbf{I}+\mathbf{X}'^*_n(\mathbf{X}'^*_m)^T+\delta_{m,n}\mathbf{S}\Big)\widehat{\mathbf{Y}}'_n\\ \overset{(a)}{\geq}&(\widehat{\mathbf{Y}}'_m)^T\Big(\Big(-\frac{1}{4}\left(\|\widehat{\mathbf{Y}}'_m\|_2^2+\|\widehat{\mathbf{Y}}'_n\|_2^2\right)-\frac{1}{\delta}\|\widehat{\mathbf{Y}}'_m\|_2^2-\delta\|\mathbf{X}'^*_n\|_2^2+(\mathbf{X}'^*_m)^T\mathbf{X}'^*_n\Big)\mathbf{I}\\ &\mathbf{X}'^*_n(\mathbf{X}'^*_m)^T+\delta_{m,n}\mathbf{S}\Big)\widehat{\mathbf{Y}}'_n\\ =&(\widehat{\mathbf{Y}}'_m)^T\left(-(\frac{1}{4}+\frac{1}{\delta})\|\widehat{\mathbf{Y}}'_m\|_2^2-\frac{1}{4}\|\widehat{\mathbf{Y}}'_n\|_2^2\right)\widehat{\mathbf{Y}}'_n\\ &+(\widehat{\mathbf{Y}}'_m)^T\Big(\left((\mathbf{X}'^*_m)^T\mathbf{X}'^*_n-\delta\|\mathbf{X}'^*_n\|_2^2\right)\mathbf{I}+\mathbf{X}'^*_n(\mathbf{X}'^*_m)^T+\delta_{m,n}\mathbf{S}\Big)\widehat{\mathbf{Y}}'_n\\ \overset{(b)}{\geq}&\|\widehat{\mathbf{Y}}'_m\|\|\widehat{\mathbf{Y}}'_n\|\left(-(\frac{1}{4}+\frac{1}{\delta})\|\widehat{\mathbf{Y}}'_m\|_2^2-\frac{1}{4}\|\widehat{\mathbf{Y}}'_n\|_2^2\right)+(\widehat{\mathbf{Y}}'_m)^T\mathcal{T}_{m,n}\widehat{\mathbf{Y}}'_n\end{aligned}$$

where

$$\mathcal{T}_{m,n}\triangleq\left((\mathbf{X}'^*_m)^T\mathbf{X}'^*_n-\delta\|\mathbf{X}'^*_n\|_2^2\right)\mathbf{I}+\mathbf{X}'^*_n(\mathbf{X}'^*_m)^T+\delta_{m,n}\mathbf{S}, \tag{A.60}$$

$\delta_{m,n}$ is the Kronecker delta function, and $\mathcal{T}_{m,n}$ is the (m,n)th block of the matrix $\mathcal{T}$, and (a) we use triangle inequality and $\delta > 0$ is any positive number; (b) we use Cauchy-Schwarz inequality.

If there exists δ such that $\mathcal{T}$ is positive definite, then $\mathbf{X}^*$ is a local minimum point of (1.1). That is, there exist some $\gamma, \epsilon > 0$ such that

$$\mathcal{L}(\mathbf{X}, \mathbf{\Omega}^*) - \mathcal{L}(\mathbf{X}^*, \mathbf{\Omega}^*) \geq \frac{\gamma}{2}\|\mathbf{X} - \mathbf{X}^*\|_F^2, \quad \forall\ \mathbf{X} \text{ such that } \|\mathbf{X}'_m - \mathbf{X}'^*_m\|_2^2 \leq \epsilon, \tag{A.61}$$

where γ is given by

$$\gamma = -\left(\frac{2K^2}{\delta} + K(K-2)\right)\epsilon^2 + 2\lambda_{\min}(\mathcal{T}) \tag{A.62}$$

where $\lambda_{\min}(\mathcal{T})$ is the smallest eigenvalue of the matrix $\mathcal{T}$. Clearly γ can be made positive for sufficiently small ϵ.

According to the definition of Lagrangian (A.50), we have

$$\mathcal{L}(\mathbf{X}, \mathbf{\Omega}^*) = f(\mathbf{X}) - \langle \mathbf{X}, \mathbf{\Omega}^* \rangle. \tag{A.63}$$

Combing with (A.61) and KKT conditions (2.4b)–(2.4d), we can obtain

$$f(\mathbf{X}) \geq \mathcal{L}(\mathbf{X}, \mathbf{\Omega}^*) \geq f(\mathbf{X}^*) + \frac{\gamma}{2}\|\mathbf{X} - \mathbf{X}^*\|_2^2, \quad \forall\ \mathbf{X} \geq 0 \text{ such that } \|\mathbf{X} - \mathbf{X}^*\| \leq \epsilon. \tag{A.64}$$

It follows that $\mathbf{X}^*$ is a strict local minimum point of problem (1.1).

A.7 Sufficient Local Optimality Condition When $K = 1$ (The proof of Corollary 1)

Proof: The term $\mathcal{M}$ is as the following.

$$\mathcal{M} = \frac{1}{2}\mathrm{Tr}[\widehat{\mathbf{Y}}[\ \widehat{\mathbf{Y}}^T \quad \mathbf{I}\ ] \begin{bmatrix} \mathbf{I} & 4\mathbf{X}^* \\ \mathbf{0} & 2(\mathbf{X}^*)^T\mathbf{X}^* \end{bmatrix} [\ \widehat{\mathbf{Y}}^T \quad \mathbf{I}\ ]^T\widehat{\mathbf{Y}}^T] + \mathrm{Tr}\left[\mathbf{X}^*\widehat{\mathbf{Y}}^T\mathbf{X}^*\widehat{\mathbf{Y}}^T\right]. \tag{A.65}$$

When $K = 1$, (A.65) becomes

$$\begin{aligned} &\frac{1}{2}\widehat{\mathbf{y}}^T\widehat{\mathbf{y}} \begin{bmatrix} \widehat{\mathbf{y}}^T & 1 \end{bmatrix} \begin{bmatrix} \mathbf{I} & 4\mathbf{x}^* \\ \mathbf{0} & 2(\mathbf{x}^*)^T\mathbf{x}^* \end{bmatrix} \begin{bmatrix} \widehat{\mathbf{y}}^T & 1 \end{bmatrix}^T + \mathrm{Tr}\left[\mathbf{x}^*\widehat{\mathbf{y}}^T\mathbf{x}^*\widehat{\mathbf{y}}^T\right] \\ =&\frac{1}{2}\widehat{\mathbf{y}}^T\widehat{\mathbf{y}}\,(\widehat{\mathbf{y}}^T\widehat{\mathbf{y}} + 4\widehat{\mathbf{y}}^T\mathbf{x}^* + 2(\mathbf{x}^*)^T\mathbf{x}^*) + \widehat{\mathbf{y}}^T\mathbf{x}^*(\mathbf{x}^*)^T\widehat{\mathbf{y}} \end{aligned} \tag{A.66}$$

where $\mathbf{x}^*$ and $\widehat{\mathbf{y}}$ denote the column of the matrix $\mathbf{X}^*$ and $\widehat{\mathbf{Y}}$.

Combining with (A.54), we have

$$\begin{aligned}\mathcal{L}(\mathbf{x},\mathbf{\Omega}^*)-\mathcal{L}(\mathbf{x}^*,\mathbf{\Omega}^*) &= \widehat{\mathbf{y}}^T\left[\frac{1}{2}\widehat{\mathbf{y}}^T\widehat{\mathbf{y}}+2\widehat{\mathbf{y}}^T\mathbf{x}^*+(\mathbf{x}^*)^T\mathbf{x}^*\right]\widehat{\mathbf{y}} \\ &\quad+\widehat{\mathbf{y}}^T\left[2\mathbf{x}^*(\mathbf{x}^*)^T-\frac{\mathbf{Z}^T+\mathbf{Z}}{2}\right]\widehat{\mathbf{y}} \\ &\overset{(a)}{\geq} \widehat{\mathbf{y}}^T\left[\frac{1}{2}\widehat{\mathbf{y}}^T\widehat{\mathbf{y}}-\frac{1}{\delta}\|\widehat{\mathbf{y}}\|_2^2-\delta\|\mathbf{x}^*\|_2^2+(\mathbf{x}^*)^T\mathbf{x}^*\right]\widehat{\mathbf{y}} \\ &\quad+\widehat{\mathbf{y}}^T\left[2\mathbf{x}^*(\mathbf{x}^*)^T-\frac{\mathbf{Z}^T+\mathbf{Z}}{2}\right]\widehat{\mathbf{y}} \\ &= \frac{1}{2}\|\widehat{\mathbf{y}}\|_2^4-\frac{1}{\delta}\|\widehat{\mathbf{y}}\|_2^4+\widehat{\mathbf{y}}^T\underbrace{\left[(1-\delta)\|\mathbf{x}^*\|_2^2\mathbf{I}+2\mathbf{x}^*(\mathbf{x}^*)^T-\frac{\mathbf{Z}^T+\mathbf{Z}}{2}\right]}_{\triangleq\mathcal{T}_1}\widehat{\mathbf{y}}\end{aligned} \tag{A.67}$$

where in (a) we have used the triangle inequality and $\delta>0$ is any positive number.

If there exists $\delta>0$ which ensures that $\mathcal{T}_1\succ 0$, then there exist some $\gamma,\epsilon>0$ such that the following is true

$$\mathcal{L}(\mathbf{x},\mathbf{\Omega}^*)-\mathcal{L}(\mathbf{x}^*,\mathbf{\Omega}^*)\geq\frac{\gamma}{2}\|\mathbf{x}-\mathbf{x}^*\|_2^2,\quad\forall\ \mathbf{x}\text{ such that }\|\mathbf{x}-\mathbf{x}^*\|\leq\epsilon. \tag{A.68}$$

In the above inequality, the constant γ is given by

$$\gamma=\left(1-\frac{2}{\delta}\right)\epsilon^2+2\lambda_{\min} \tag{A.69}$$

where $\lambda_{\min}(\mathcal{T}_1)$ denotes the smallest eigenvalue of $\mathcal{T}_1$. Clearly γ can be made positive by setting ϵ sufficiently small.

According to the definition of Lagrangian, we have

$$\mathcal{L}(\mathbf{x},\mathbf{\Omega}^*)=f(\mathbf{x})-\langle\mathbf{x},\mathbf{\Omega}^*\rangle. \tag{A.70}$$

Therefore, combining with (A.68) and the KKT conditions, we can obtain

$$f(\mathbf{x})\geq\mathcal{L}(\mathbf{x},\mathbf{\Omega}^*)\geq f(\mathbf{x}^*)+\frac{\gamma}{2}\|\mathbf{x}-\mathbf{x}^*\|_2^2.\quad\forall\ \mathbf{x}\geq 0\text{ such that }\|\mathbf{x}-\mathbf{x}^*\|\leq\epsilon. \tag{A.71}$$

APPENDIX B. PROOFS OF PA-GD

B.1 Proofs of the Preliminary Lemmas

We provide the proofs of some preliminary lemmas (Lemma 11–Lemma 13) used in the proof of Section B.2.

First, Lemma 11 and Lemma 12 give the property that quantify the size of the difference of the second-order information of the objective values between two points.

Lemma 11. *If function $f(\cdot)$ is ρ-Hessian Lipschitz, we have*

$$\left\| \int_0^1 \nabla^2 f(\theta \mathbf{x}) d\theta - \nabla^2 f(\mathbf{y}) \right\| \leq \rho \left(\|\mathbf{x}\| + \|\mathbf{y}\| \right), \quad \forall \mathbf{x}, \mathbf{y}. \tag{B.1}$$

Lemma 12. *Under Assumption 1, we have block-wise Lipschitz continuity as follows:*

$$\left\| \begin{bmatrix} \nabla_{11}^2 f(\mathbf{x}) & \nabla_{12}^2 f(\mathbf{x}) \\ \mathbf{0} & \nabla_{22}^2 f(\mathbf{y}) \end{bmatrix} - \begin{bmatrix} \nabla_{11}^2 f(\mathbf{z}) & \nabla_{12}^2 f(\mathbf{z}) \\ \mathbf{0} & \nabla_{22}^2 f(\mathbf{z}) \end{bmatrix} \right\| \leq \rho \left(\|\mathbf{x} - \mathbf{z}\| + \|\mathbf{y} - \mathbf{z}\| \right), \forall \mathbf{x}, \mathbf{y}, \mathbf{z}, \tag{B.2}$$

and

$$\left\| \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \nabla_{21}^2 f(\mathbf{x}) & \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \nabla_{21}^2 f(\mathbf{y}) & \mathbf{0} \end{bmatrix} \right\| \leq \rho \|\mathbf{x} - \mathbf{y}\|, \forall \mathbf{x}, \mathbf{y}. \tag{B.3}$$

Then, we illustrate that the size of the partial gradient with one round update by the AGD algorithm has the following relation with the full size of the gradient.

Lemma 13. *If function $f(\cdot)$ is L-smooth with Lipschitz constant, then we have*

$$\|\nabla f(\mathbf{x}^{(t)})\|^2 \leq 4 \sum_{k=1}^{2} \|\nabla_k f(\mathbf{h}_{-k}^{(t)}, \mathbf{x}_k^{(t)})\|^2 \tag{B.4}$$

where sequence $\mathbf{x}_k^{(t)}, k = 1, 2$ is generated by the AGD algorithm.

B.1.1 Proof of Lemma 11

Proof. If function $f(\cdot)$ is ρ-Hessian Lipschitz, then we have

$$\begin{aligned}&\left\|\int_0^1 (\nabla^2 f(\theta\mathbf{x}) - \nabla^2 f(\mathbf{y}))d\theta\right\| \leq \int_0^1 \left\|\nabla^2 f(\theta\mathbf{x}) - \nabla^2 f(\mathbf{y})\right\| d\theta \\ \overset{(a)}{\leq} &\rho \int_0^1 \|\theta\mathbf{x} - \mathbf{y}\|\, d\theta \overset{(b)}{\leq} \rho \int_0^1 \theta\|\mathbf{x}\| d\theta + \rho\|\mathbf{y}\| \leq \rho\left(\|\mathbf{x}\| + \|\mathbf{y}\|\right)\end{aligned}$$

where (a) is true because of Hessian Lipschitz, in (b) we used the triangle inequality. □

B.1.2 Proof of Lemma 12

There proof involves two parts:

Upper Triangular Matrix: Consider three different vectors $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$. We can have

$$\begin{aligned}&\left\|\begin{bmatrix} \nabla^2_{11} f(\mathbf{x}) & \nabla^2_{12} f(\mathbf{x}) \\ 0 & \nabla^2_{22} f(\mathbf{y}) \end{bmatrix} - \begin{bmatrix} \nabla^2_{11} f(\mathbf{z}) & \nabla^2_{12} f(\mathbf{z}) \\ 0 & \nabla^2_{22} f(\mathbf{z}) \end{bmatrix}\right\| \\ \leq &\left\|\mathbf{I}_1 \left(\begin{bmatrix} \nabla^2_{11} f(\mathbf{x}) & \nabla^2_{12} f(\mathbf{x}) \\ \nabla^2_{21} f(\mathbf{x}) & \nabla^2_{22} f(\mathbf{x}) \end{bmatrix} - \begin{bmatrix} \nabla^2_{11} f(\mathbf{z}) & \nabla^2_{12} f(\mathbf{z}) \\ \nabla^2_{21} f(\mathbf{z}) & \nabla^2_{22} f(\mathbf{z}) \end{bmatrix} \right)\right\| \\ &+ \left\|\mathbf{I}_2 \left(\begin{bmatrix} \nabla^2_{11} f(\mathbf{y}) & \nabla^2_{12} f(\mathbf{y}) \\ \nabla^2_{21} f(\mathbf{y}) & \nabla^2_{22} f(\mathbf{y}) \end{bmatrix} - \begin{bmatrix} \nabla^2_{11} f(\mathbf{z}) & \nabla^2_{12} f(\mathbf{z}) \\ \nabla^2_{21} f(\mathbf{z}) & \nabla^2_{22} f(\mathbf{z}) \end{bmatrix} \right) \mathbf{I}_2\right\| \\ \overset{(a)}{\leq} &\left\|\begin{bmatrix} \nabla^2_{11} f(\mathbf{x}) & \nabla^2_{12} f(\mathbf{x}) \\ \nabla^2_{21} f(\mathbf{x}) & \nabla^2_{22} f(\mathbf{x}) \end{bmatrix} - \begin{bmatrix} \nabla^2_{11} f(\mathbf{z}) & \nabla^2_{12} f(\mathbf{z}) \\ \nabla^2_{21} f(\mathbf{z}) & \nabla^2_{22} f(\mathbf{z}) \end{bmatrix}\right\| \\ &+ \left\|\begin{bmatrix} \nabla^2_{11} f(\mathbf{y}) & \nabla^2_{12} f(\mathbf{y}) \\ \nabla^2_{21} f(\mathbf{y}) & \nabla^2_{22} f(\mathbf{y}) \end{bmatrix} - \begin{bmatrix} \nabla^2_{11} f(\mathbf{z}) & \nabla^2_{12} f(\mathbf{z}) \\ \nabla^2_{21} f(\mathbf{z}) & \nabla^2_{22} f(\mathbf{z}) \end{bmatrix}\right\| \\ \leq &\rho\left(\|\mathbf{x} - \mathbf{z}\| + \|\mathbf{y} - \mathbf{z}\|\right)\end{aligned}$$

where in (a) we used

$$\mathbf{I}_1 = \begin{bmatrix} \mathbf{I} & 0 \\ 0 & 0 \end{bmatrix} \qquad \mathbf{I}_2 = \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{I} \end{bmatrix} \tag{B.5}$$

and $\|\mathbf{I}_1\| = \|\mathbf{I}_2\| = 1$.

Lower Triangular Matrix:

$$\begin{aligned}
&\left\| \begin{bmatrix} 0 & 0 \\ \nabla^2_{21} f(\mathbf{x}) & 0 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ \nabla^2_{21} f(\mathbf{y}) & 0 \end{bmatrix} \right\| \\
&= \left\| \mathbf{I}_2 \left(\begin{bmatrix} \nabla^2_{11} f(\mathbf{x}) & \nabla^2_{12} f(\mathbf{x}) \\ \nabla^2_{21} f(\mathbf{x}) & \nabla^2_{22} f(\mathbf{x}) \end{bmatrix} - \begin{bmatrix} \nabla^2_{11} f(\mathbf{y}) & \nabla^2_{12} f(\mathbf{y}) \\ \nabla^2_{21} f(\mathbf{y}) & \nabla^2_{22} f(\mathbf{y}) \end{bmatrix} \right) \mathbf{I}_1 \right\| \\
&\overset{(a)}{\leq} \rho \|\mathbf{x} - \mathbf{y}\|
\end{aligned}$$

where (a) is true because we know $\|\mathbf{I}_1\| = \|\mathbf{I}_2\| = 1$.

B.1.3 Proof of Lemma 13

Proof. Recall the definition

$$\mathbf{h}^{(t)}_{-1} := \mathbf{x}^{(t)}_2 \quad \text{and} \quad \mathbf{h}^{(t)}_{-2} := \mathbf{x}^{(t+1)}_1.$$

First, we have

$$\|\nabla_2 f(\mathbf{x}^{(t)}_1, \mathbf{x}^{(t)}_2)\|^2 \leq 2\|\nabla_2 f(\mathbf{x}^{(t+1)}_1, \mathbf{x}^{(t)}_2) - \nabla_2 f(\mathbf{x}^{(t)}_1, \mathbf{x}^{(t)}_2)\|^2 + 2\|\nabla_2 f(\mathbf{x}^{(t+1)}_1, \mathbf{x}^{(t)}_2)\|^2. \tag{B.6}$$

Using block-wise Lipschitz continuity, we have

$$\begin{aligned}
\|\nabla_2 f(\mathbf{x}^{(t)}_1, \mathbf{x}^{(t)}_2)\|^2 &\leq 2L^2_{\max}\|\mathbf{x}^{(t+1)}_1 - \mathbf{x}^{(t)}_1\|^2 + 2\|\nabla_2 f(\mathbf{x}^{(t+1)}_1, \mathbf{x}^{(t)}_2)\|^2 \\
&\overset{(a)}{=} 2L^2_{\max}\|\eta \nabla_1 f(\mathbf{x}^{(t)}_1, \mathbf{x}^{(t)}_2)\|^2 + 2\|\nabla_2 f(\mathbf{x}^{(t+1)}_1, \mathbf{x}^{(t)}_2)\|^2 \\
&\overset{(b)}{\leq} 2\sum_{k=1}^{2} \|\nabla_k f(\mathbf{h}^{(t)}_{-k}, \mathbf{x}^{(t)}_k)\|^2
\end{aligned} \tag{B.7}$$

where (a) is because we use the update rule of AGD, (b) is true due to $\eta \leq 1/L_{\max}$.

Summing $\|\nabla_1 f(\mathbf{x}^{(t)}_1, \mathbf{x}^{(t)}_2)\|^2$ on both sides of the above equation, we have

$$\|\nabla f(\mathbf{x}^{(t)})\|^2 \leq \sum_{k=1}^{2} \|\nabla_k f(\mathbf{x}^{(t)}_k)\|^2 \leq 4\sum_{k=1}^{2} \|\nabla_k f(\mathbf{h}^{(t)}_{-k}, \mathbf{x}^{(t)}_k)\|^2. \tag{B.8}$$

□

B.2 Proofs of the Convergence Rate of PA-GD

As stated in the main body of the dissertation, we can use Lemma 5 and Lemma 6 to prove Theorem 8. Lemma 5 is basically well-known. The main task focuses on proving Lemma 6, which consists of a sequence of lemmas (Lemma 14–Lemma 16) that lead to Lemma 6.

Before discussing the details of Lemma 6, we need to introduce some constants defined as follows,

$$\begin{aligned}
\mathcal{F} :=& \eta^5 L_{\max}^5 \frac{\gamma^3}{\kappa^3 \rho^2} \log^{-6}\left(\frac{d\kappa}{\delta}\right) \mathcal{P}_1^{-6} \mathcal{P}_2^{-2}, \\
\mathcal{G} :=& \eta^2 L_{\max}^2 \frac{\gamma^2}{\rho} \log^{-3}\left(\frac{d\kappa}{\delta}\right) \mathcal{P}_1^{-3} \mathcal{P}_2^{-1}, \\
\mathcal{S} :=& \eta^2 L_{\max}^2 \frac{\gamma}{\kappa\rho} \log^{-2}\left(\frac{d\kappa}{\delta}\right) \mathcal{P}_1^{-2} \mathcal{P}_2^{-1}, \\
\mathcal{T} :=& \frac{\log\left(\frac{d\kappa}{\delta}\right) \mathcal{P}_1}{\eta\gamma}.
\end{aligned}$$

These quantities refer to different units of the algorithm. Specifically, $\mathcal{F}$ accounts for the objective value, $\mathcal{G}$ for the size of the gradient, $\mathcal{S}$ for the norm of the difference between iterates, and $\mathcal{T}$ for the number of iterations. Also, we define a condition number in terms of γ as $\kappa := \frac{L_{\max}}{\gamma} \geq 1$.

These quantities, $\mathcal{F}$, $\mathcal{G}$, $\mathcal{S}$ and $\mathcal{T}$ have some certain relations as follows, which are useful of simplifying the expressions in the proofs.

$$\sqrt{\mathcal{F}} = \frac{\sqrt{\eta}\mathcal{G}}{\kappa}, \tag{B.9a}$$

$$\frac{\eta\mathcal{G}\mathcal{T}}{\kappa} = \mathcal{S}, \tag{B.9b}$$

$$\rho\mathcal{S}^3 = \frac{\eta L_{\max}\mathcal{F}}{\mathcal{P}_2}, \tag{B.9c}$$

$$\eta\rho\mathcal{S}\mathcal{T} = \frac{\eta^2 L_{\max}^2}{\kappa \log(\frac{d\kappa}{\delta})\mathcal{P}_1\mathcal{P}_2}. \tag{B.9d}$$

In the process of the proofs, we used conditions $\log(\frac{d\kappa}{\delta}) \geq 1$, $\mathcal{P}_1 \geq 2$ repeatedly to simply the expressions of the parameters. We also consider saddle point $\widetilde{\mathbf{x}}^{(t)}$ that satisfies the following condition.

Condition 1. *An ϵ-second order stationary point $\widetilde{\mathbf{x}}^{(t)}$ satisfies the following conditions:*

$$\sum_{k=1}^{2} \|\nabla_k f(\widetilde{\mathbf{h}}_{-k}^{(t)}, \widetilde{\mathbf{x}}_k^{(t)})\|^2 \leq g_{th}^2 \quad \textit{and} \quad \lambda_{\min}(\nabla^2 f(\widetilde{\mathbf{x}}^{(t)})) \leq -\gamma \tag{B.10}$$

where $g_{th} := \frac{\mathcal{G}}{2\kappa}$.

Condition 1 implies that point $\widetilde{\mathbf{x}}^{(t)}$ satisfies $\|\nabla f(\widetilde{\mathbf{x}}^{(t)})\| \leq \mathcal{G}/\kappa$ (see Lemma 13) and $\lambda_{\min}(\nabla^2 f(\widetilde{\mathbf{x}}^{(t)})) \leq -\gamma$.

Sufficient Decrease after Perturbation Consider $\widetilde{\mathbf{x}}^{(t)}$ satisfy Condition 1 and let $\mathbf{H} \triangleq \nabla^2 f(\widetilde{\mathbf{x}}^{(t)})$. We consider a second order approximation as the following

$$\widehat{f}_{\mathbf{y}}(\mathbf{x}) \triangleq f(\mathbf{y}) + \nabla f(\mathbf{y})^T(\mathbf{x}-\mathbf{y}) + \frac{1}{2}(\mathbf{x}-\mathbf{y})^T\mathbf{H}(\mathbf{x}-\mathbf{y}). \tag{B.11}$$

With these definitions of parameters, we will study how PA-GD can escape from strict saddle points. The main part of the proof is to show that when two sequences are apart from each other with a certain distance along the $\vec{\mathbf{e}}$ direction at the starting points, where $\vec{\mathbf{e}}$ denotes the eigenvector of $\mathbf{M}^{-1}\mathbf{T}$ whose eigenvalue is maximum (greater than 1). Then, after a number of iterations at least one of them can give a sufficient decrease of the objective value. This property implies the iterates can easily escape from the saddle points as long as there is a large enough perturbation between the initial points of the two sequences along the $\vec{\mathbf{e}}$ direction. We will introduce the following two lemmas formally which are the main contributions of this work.

Lemma 14. *Under Assumption 1, consider $\widetilde{\mathbf{x}}^{(t)}$ that satisfies Condition 1 and a generic sequence $\mathbf{u}^{(t)}$ generated by AGD. For any constant $\widehat{c} \geq 2$, $\delta \in (0, \frac{d\kappa}{e}]$, when initial point $\mathbf{u}^{(0)}$ satisfies*

$$\|\mathbf{u}^{(0)} - \widetilde{\mathbf{x}}^{(t)}\| \leq 2r, \tag{B.12}$$

then, with the definition of

$$r := \frac{\eta L_{\max}\mathcal{S}}{\kappa \log(\frac{d\kappa}{\delta})\mathcal{P}_1}, \quad \textit{and} \quad T := \min\{\inf_t\{t | \widehat{f}_{\mathbf{u}^{(0)}}(\mathbf{u}^{(t)}) - f(\mathbf{u}^{(0)}) \leq -3\mathcal{F}\}, \widehat{c}\mathcal{T}\}, \tag{B.13}$$

there exits constants $c_{\max}^{(1)}, \widehat{c}$ such that for any $\eta \leq c_{\max}^{(1)}/L_{\max}$, the iterates generated by PA-GD satisfy $\|\mathbf{u}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\| \leq 5\widehat{c}\mathcal{S}, \forall t < T$.

Lemma 15. *Under Assumption 1, consider $\widetilde{\mathbf{x}}^{(t)}$ that satisfies Condition 1. There exist constants $c_{\max}^{(2)}$, $\widehat{c}$ such that: for any $\delta \in (0, \frac{d\kappa}{e}]$ and $\eta \leq c_{\max}^{(2)}/L_{\max}$, with the definition of*

$$T := \min\left\{\inf_t\{t|\widehat{f}_{\mathbf{w}_0}(\mathbf{w}^{(t)}) - f(\mathbf{w}^{(0)}) \leq -3\mathcal{F}\}, \widehat{c}\mathcal{T}\right\}$$

where two iterates $\{\mathbf{u}^{(t)}\}$ and $\{\mathbf{w}^{(t)}\}$ that are generated by PA-GD with initial points $\{\mathbf{u}^{(0)}, \mathbf{w}^{(0)}\}$ satisfying

$$\|\mathbf{u}^{(0)} - \widetilde{\mathbf{x}}^{(t)}\| \leq r,\ \mathbf{w}^{(0)} = \mathbf{u}^{(0)} + \upsilon r\vec{\mathbf{e}},\ \upsilon \in [\delta/(2\sqrt{d}), 1], \tag{B.14}$$

where $\vec{\mathbf{e}}$ denotes the eigenvector of $\mathbf{M}^{-1}\mathbf{T}$ whose eigenvalue is maximum, then, if $\|\mathbf{u}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\| \leq 5\widehat{c}\mathcal{S}, \forall t < T$, we will have $T < \widehat{c}\mathcal{T}$.

Lemma 14 says that if the $\mathbf{u}^{(t)}$-iterate generated by PA-GD cannot provide a sufficient decrease of the objective value, then the iterates are constrained within the area which is very close to the saddle point. With this property, Lemma 15 shows if there exists another PA-GD iterate $\mathbf{w}^{(t)}$, which is initialized with a certain distance along the $\vec{\mathbf{e}}$ direction from the $\mathbf{u}$-iterate, then $\mathbf{w}^{(t)}$ will provide a sufficient decrease of the objective value. These two lemmas characterize the convergence behavior of the PA-GD iterates.

Escaping from Saddle Points Then, we need to quantify the probability that after adding the perturbation the algorithm cannot escape from strict saddle points. In previous work about escaping from saddle points with GD, a characterization of the geometry around saddle points has been given (101, Lemma 15). Once we know that PA-GD also decreases the objective value sufficiently in Lemma 14 and Lemma 15, the following lemma can be claimed straightforwardly. To be more specific, we can obtain the probability that iterates will be stuck at the strict points after T iterations as follows.

$$\begin{aligned}
\mathbb{P}(\mathbf{w}^{(0)} \in \mathcal{X}_{\text{stuck}}) &= \int_{\mathbb{B}_{\widetilde{\mathbf{x}}^{(t)}}(r)} \mathbb{P}(\mathbf{w}^{(0)} \in \mathcal{X}_{\text{stuck}}|\mathbf{u}^{(0)} \in \mathcal{X}_{\text{stuck}})\mathbb{P}(\mathbf{u}^{(0)} \in \mathcal{X}_{\text{stuck}})d\mathbf{u}^{(0)} \\
&\leq \int_{\mathbb{B}_{\widetilde{\mathbf{x}}^{(t)}}(r)} \mathbb{P}(\mathbf{w}^{(0)} \in \mathcal{X}_{\text{stuck}}|\mathbf{u}^{(0)} \in \mathcal{X}_{\text{stuck}})\mathbb{P}(\mathbf{u}^{(0)})d\mathbf{u}^{(0)} \\
&\overset{(a)}{\leq} \delta \int_{\mathbb{B}_{\widetilde{\mathbf{x}}^{(t)}}(r)} \mathbb{P}(\mathbf{u}^{(0)})d\mathbf{u}^{(0)} = \delta
\end{aligned}$$

where $\mathcal{X}_{\text{stuck}}$ denotes the set where the algorithm starts such that the sequence cannot escape from the strict saddle point after T iterations, (a) is true because probability $\mathbb{P}(\mathbf{w}^{(0)} \in \mathcal{X}_{\text{stuck}} | \mathbf{u}^{(0)} \in \mathcal{X}_{\text{stuck}})$ can be upper bounded by δ, which is proven in the following lemma.

Lemma 16. *Under Assumption 1, there exists a universal constant $c_{\max}$, for any $\delta \in (0, d\kappa/e]$: consider a saddle point $\widetilde{\mathbf{x}}^{(t)}$ which satisfies Condition 1, let $\mathbf{x}^{(0)} = \widetilde{\mathbf{x}}^{(t)} + \xi$ where ξ is generated randomly which follows the uniform distribution over a ball with radius r, and let $\mathbf{x}^{(t)}$ be the iterates of PA-GD starting from $\mathbf{x}^{(0)}$. Then, when step size $\eta \leq c_{\max}/L_{\max}$, with at least probability $1-\delta$, we have the following for any $T \geq \mathcal{T}/c_{\max}$*

$$f(\mathbf{x}^{(T)}) - f(\widetilde{\mathbf{x}}^{(t)}) \leq -\mathcal{F}. \tag{B.15}$$

Then, applying $\eta = \frac{c}{L_{\max}}$,$\gamma = (L_{\max}\rho\epsilon)^{1/3}$, and $\delta = \frac{dL_{\max}}{(L_{\max}\rho\epsilon)^{1/3}} e^{-\chi}$ into Lemma 16, we can get Lemma 6 immediately.

With these lemmas, we can give the proof of Theorem 8 as the following.

B.2.1 Proof of Theorem 8

Next, we prove the main theorem.

Proof. Submitting $\eta = \frac{c}{L_{\max}}$,$\gamma = (L_{\max}\rho\epsilon)^{1/3}$, and $\delta = \frac{dL_{\max}}{(L_{\max}\rho\epsilon)^{1/3}} e^{-\chi}$ into the definitions of $\mathcal{F}, \mathcal{G}, \mathcal{T}$, we will have the following definitions.

$$\begin{aligned}
f_{\mathsf{th}} :=& \mathcal{F} = \frac{c^5\epsilon^2}{L_{\max}(\chi\mathcal{P}_1)^6\mathcal{P}_2^2}, \\
g_{\mathsf{th}} :=& \frac{\mathcal{G}}{2\kappa} = \frac{c^2\epsilon}{2(\chi\mathcal{P}_1)^3\mathcal{P}_2}, \\
t_{\mathsf{th}} :=& \frac{\mathcal{T}}{c} = \frac{L_{\max}\chi\mathcal{P}_1}{c^2(L_{\max}\rho\epsilon)^{\frac{1}{3}}}.
\end{aligned}$$

After applying Lemma 13, we know that

$$\|\nabla f(\mathbf{x})\| \leq \frac{c}{\chi^3\mathcal{P}_1^3\mathcal{P}_2}\epsilon$$

where $c \leq 1, \chi, \mathcal{P}_1, \mathcal{P}_2 \geq 1$.

With a set of necessary lemmas and leveraging the proof of PGD (101, Theorem 3), we have the following convergence analysis of PA-GD. Specifically, at any iteration, we need to consider two cases (we use the first iteration as an example):

1. In this case the gradient is large such that $\sum_{k=1}^{2} \|\nabla_k f(\mathbf{h}_{-k}^{(0)}, \mathbf{x}_k^{(0)})\|^2 > g_{\mathsf{th}}^2$: According to Lemma 5, we have

$$f(\mathbf{x}^{(1)}) - f(\mathbf{x}^{(0)}) \leq -\sum_{k=1}^{2} \frac{\eta}{2} \|\nabla_k f(\mathbf{h}_{-k}^{(0)}, \mathbf{x}_k^{(0)})\|^2 \leq -\frac{\eta}{2} g_{\mathsf{th}}^2$$
$$\overset{(a)}{=} -\frac{c^5}{8(\chi \mathcal{P}_1)^6 \mathcal{P}_2^2} \frac{\epsilon^2}{L_{\max}} \tag{B.16}$$

   where in (a) use the definition of g_{th}^2 and $\eta \leq c/L_{\max}$.

2. The gradient is small in all block directions, namely $\sum_{k=1}^{2} \|\nabla_k f(\mathbf{h}_{-k}^{(0)}, \mathbf{x}_k^{(0)})\|^2 \leq g_{\mathsf{th}}^2$: in this case, we will add the perturbation to the iterates, and implement AGD for the next t_{th} steps and then check the termination condition. If the termination condition is not satisfied, we must have

$$f(\mathbf{x}^{(t_{\mathsf{th}})}) - f(\mathbf{x}^{(0)}) \leq -f_{\mathsf{th}} = -\frac{c^5 \epsilon^2}{L_{\max}(\chi \mathcal{P}_1)^6 \mathcal{P}_2^2}, \tag{B.17}$$

   which implies that the objective value in each step on average is decreased by

$$\frac{f(\mathbf{x}^{(t_{\mathsf{th}})}) - f(\mathbf{x}^{(0)})}{t_{\mathsf{th}}} \leq -\frac{c^7}{(\chi \mathcal{P}_1)^7 \mathcal{P}_2^2} \frac{\epsilon^2}{L_{\max}} \frac{(L_{\max}\rho\epsilon)^{\frac{1}{3}}}{L_{\max}}. \tag{B.18}$$

   Since $\kappa = L_{\max}/(L_{\max}\rho\epsilon)^{1/3} \geq 1$, we know that the right-hand side (RHS) of (B.18) is greater than RHS of (B.16).

   With the results of these two cases, we can know that if there is a large size of the gradient, we can know the decrease of the objective function value by the result of case 1, and if not, we use the result of case 2. In summary, PA-GD can have a sufficient decrease of the objective function value by $\frac{c^7}{(\chi \mathcal{P}_1)^7 \mathcal{P}_2^2} \frac{\epsilon^2}{L_{\max}} \frac{(L_{\max}\rho\epsilon)^{1/3}}{L_{\max}}$ per iteration on average. This means that Algorithm 2 must stop within a finite number of iterations, which is

$$\frac{f(\mathbf{h}_{-1}^{(0)}, \mathbf{x}_1^{(0)}) - f^*}{\frac{c^7}{(\chi \mathcal{P}_1)^7 \mathcal{P}_2^2} \frac{\epsilon^2}{L_{\max}} \frac{(L_{\max}\rho\epsilon)^{1/3}}{L_{\max}}} = \frac{(\chi \mathcal{P}_1)^7 \mathcal{P}_2^2}{c^7} \frac{L_{\max}^2 \Delta f}{\epsilon^2 (L_{\max}\rho\epsilon)^{1/3}} = \mathcal{O}\left(\frac{\Delta f (\chi \mathcal{P}_1)^7 \mathcal{P}_2^2 L_{\max}^{5/3}}{\rho^{1/3} \epsilon^{7/3}}\right) \tag{B.19}$$

where $\Delta f := f(\mathbf{h}_{-1}^{(0)}, \mathbf{x}_1^{(0)}) - f^*$.

According to Lemma 6, we know that with probability $1 - \frac{dL_{\max}}{(L_{\max}\rho\epsilon)^{1/3}} e^{-\chi}$ the algorithm can give a sufficient descent with the perturbation when $\sum_{k=1}^{2} \|\nabla_k f(\mathbf{h}_{-k}^{(t)}, \mathbf{x}_k^{(t)})\|^2 \leq g_{\mathsf{th}}^2$. Since the total number of perturbation we can add is at most

$$n = \frac{1}{t_{\mathsf{th}}} \frac{(\chi \mathcal{P}_1)^7 \mathcal{P}_2^2}{c^7} \frac{L_{\max}^2 \Delta f}{\epsilon^2 (L_{\max}\rho\epsilon)^{1/3}} = \frac{(\mathcal{P}_1\chi)^6 \mathcal{P}_2^2}{c^5} \frac{L_{\max}\Delta_f}{\epsilon^2}. \tag{B.20}$$

Using the union bound, the probability of Lemma 6 being satisfied for all perturbations is

$$\begin{aligned} 1 - n\frac{dL_{\max}}{(L_{\max}\rho\epsilon)^{\frac{1}{3}}} e^{-\chi} &= 1 - \frac{dL_{\max}}{(L_{\max}\rho\epsilon)^{\frac{1}{3}}} e^{-\chi} \frac{(\mathcal{P}_1\chi)^6 \mathcal{P}_2^2}{c^5} \frac{L_{\max}\Delta_f}{\epsilon^2} \\ &= 1 - \underbrace{\frac{dL_{\max}}{(L_{\max}\rho\epsilon)^{\frac{1}{3}}} \frac{\mathcal{P}_1^6 \mathcal{P}_2^2}{c^5} \frac{\Delta_f}{\epsilon^2}}_{:=\mathcal{C}} \chi^6 e^{-\chi}. \end{aligned} \tag{B.21}$$

With chosen $\chi = 6\max\{\ln(\mathcal{C}/\delta), 4\}$, we have $\chi^6 e^{-\chi} \leq e^{-\chi/6}$, which implies $\chi^6 e^{-\chi}\mathcal{C} \leq e^{-\chi/6}\mathcal{C} \leq \delta$.

The proof is complete. □

B.2.2 Proof of Lemma 4

Proof. Recall the definitions:

$$\mathbf{H}_u := \begin{bmatrix} \nabla_{11}^2 f(\widetilde{\mathbf{x}}^{(t)}) & \nabla_{12}^2 f(\widetilde{\mathbf{x}}^{(t)}) \\ 0 & \nabla_{22}^2 f(\widetilde{\mathbf{x}}^{(t)}) \end{bmatrix} \quad \mathbf{H}_l := \begin{bmatrix} 0 & 0 \\ \nabla_{21}^2 f(\widetilde{\mathbf{x}}^{(t)}) & 0 \end{bmatrix}, \tag{B.22}$$

where $\widetilde{\mathbf{x}}^{(t)}$ is an ϵ-second order stationary point, and

$$\mathbf{M} := \mathbf{I} + \eta \mathbf{H}_l, \quad \mathbf{T} := \mathbf{I} - \eta \mathbf{H}_u. \tag{B.23}$$

Our goal of this lemma is to show that the maximum eigenvalue of $\mathbf{M}^{-1}\mathbf{T}$ is greater than 1 so that we can project iterates $\mathbf{v}^{(t)}$ onto the two subspaces, where the first subspace is spanned by the eigenvector of $\mathbf{M}^{-1}\mathbf{T}$ whose eigenvalue is the largest (greater than 1) and the other one is spanned by the remaining eigenvectors.

Note that $\det(\mathbf{M}) = 1$, which implies that $\det(\mathbf{M}^{-1}\mathbf{T} - \lambda\mathbf{I}) = \det(\mathbf{T} - \lambda\mathbf{M})$, where λ denotes the eigenvalue. We can analyze the determinant of $\mathbf{T} - \lambda\mathbf{M}$, i.e.,

$$\begin{aligned}\det[\mathbf{T} - \lambda\mathbf{M}] =& \det[\mathbf{I} - \eta\mathbf{H}_u - \lambda(\mathbf{I} + \eta\mathbf{H}_l)] \\ =& \det\left[\underbrace{\begin{array}{cc} (1-\lambda)\mathbf{I} - \eta\nabla_{11}^2 f(\widetilde{\mathbf{x}}^{(t)}) & -\eta\nabla_{12}^2 f(\widetilde{\mathbf{x}}^{(t)}) \\ -\lambda\eta\nabla_{21}^2 f(\widetilde{\mathbf{x}}^{(t)}) & (1-\lambda)\mathbf{I} - \eta\nabla_{22}^2 f(\widetilde{\mathbf{x}}^{(t)}) \end{array}}_{:=\mathbf{Q}(\lambda)}\right].\end{aligned}$$

Then, we use two steps to show $\lambda_{\max}(\mathbf{M}^{-1}\mathbf{T}) > 1$: 1) we can show that all eigenvalues of $\mathbf{Q}(\lambda)$ are real; 2) there exists a $\lambda > 1$ such that $\det(\mathbf{Q}(\lambda)) = 0$.

Consider a $\delta > 0$. We have

$$\mathbf{Q}(1+\delta) = -\left(\underbrace{\eta\mathbf{H} + \delta(\mathbf{I} + \eta\mathbf{H}_l)}_{:=\mathbf{F}(\delta)}\right) \tag{B.24}$$

where

$$\begin{aligned}\mathbf{F}(\delta) =& \delta\mathbf{I} + \eta\begin{bmatrix} \nabla_{11}^2 f(\widetilde{\mathbf{x}}^{(t)}) & \nabla_{12}^2 f(\widetilde{\mathbf{x}}^{(t)}) \\ (1+\delta)\nabla_{21}^2 f(\widetilde{\mathbf{x}}^{(t)}) & \nabla_{22}^2 f(\widetilde{\mathbf{x}}^{(t)}) \end{bmatrix} \\ =& \begin{bmatrix} \mathbf{I} & \\ & \sqrt{1+\delta} \end{bmatrix} \underbrace{\begin{bmatrix} \delta\mathbf{I} + \eta\nabla_{11}^2 f(\widetilde{\mathbf{x}}^{(t)}) & \eta\sqrt{1+\delta}\nabla_{12}^2 f(\widetilde{\mathbf{x}}^{(t)}) \\ \eta\sqrt{1+\delta}\nabla_{21}^2 f(\widetilde{\mathbf{x}}^{(t)}) & \delta\mathbf{I} + \eta\nabla_{22}^2 f(\widetilde{\mathbf{x}}^{(t)}) \end{bmatrix}}_{\mathbf{G}(\delta)} \begin{bmatrix} \mathbf{I} & \\ & \frac{1}{\sqrt{1+\delta}} \end{bmatrix},\end{aligned}$$

meaning that $\mathbf{F}(\delta)$ is similar to $\mathbf{G}(\delta)$. Consequently, we can conclude that $\mathbf{F}(\delta)$ has the same eigenvalues of $\mathbf{G}(\delta)$. Since we know that $\mathbf{H}$ and $\mathbf{G}(\delta)$ are diagonalizable (normal matrices), then we have the following result (107) (or (108)) of quantifying the difference of the eigenvalues of the two normal matrices

$$\max_{1\le i\le d} |\lambda_i(\eta\mathbf{H}) - \lambda_i(\mathbf{G}(\delta))| \le \|\eta\mathbf{H} - \mathbf{G}(\delta)\| \tag{B.25}$$

where $\lambda_i(\mathbf{H})$ and $\lambda_i(\mathbf{G}(\delta))$ denote the ith eigenvalue of $\mathbf{H}$ and $\mathbf{G}(\delta)$, which are listed in a decreasing order.

With the help of (B.25), we can check

$$\begin{aligned}
&\|\eta\mathbf{H}-\mathbf{G}(\delta)\| \\
=&\left\|\delta\mathbf{I}+\begin{bmatrix} 0 & (\sqrt{1+\delta}-1)\eta\nabla^2_{12}f(\widetilde{\mathbf{x}}^{(t)}) \\ (\sqrt{1+\delta}-1)\eta\nabla^2_{21}f(\widetilde{\mathbf{x}}^{(t)}) & 0 \end{bmatrix}\right\| \\
\leq&\delta+(\sqrt{1+\delta}-1)\eta\|\mathbf{H}\|+(\sqrt{1+\delta}-1)\eta\left\|\begin{matrix} \nabla^2_{11}f(\widetilde{\mathbf{x}}^{(t)}) & 0 \\ 0 & \nabla^2_{22}f(\widetilde{\mathbf{x}}^{(t)}) \end{matrix}\right\| \\
\overset{(a)}{\leq}&\delta+(\sqrt{1+\delta}-1)(\frac{L}{L_{\max}}+1)
\end{aligned} \tag{B.26}$$

where (a) is true since we used $\eta \leq c_{\max}/L_{\max}$ and the fact that $\|\mathbf{H}\| \leq L$ and $\|\mathbf{H}_d\| \leq L_{\max}$. Also, it can be observed that when $\delta = 0$, matrix $\mathbf{G}(\delta)$ is reduced to $\eta\mathbf{H}$. Note that if $\eta = 1/L$ is used, then we have $\|\eta\mathbf{H}-\mathbf{G}(\delta)\| \leq \delta + 2(\sqrt{1+\delta}-1)$.

We know that the minimum eigenvalue of $\eta\mathbf{H}$ which is $-\eta\gamma$ and the maximum difference of the eigenvalues between $\eta\mathbf{H}$ and $\mathbf{G}(\delta)$ is upper bounded by (B.26). Then, we can choose a sufficient small δ such that $\mathbf{G}(\delta)$ also has a negative eigenvalue, meaning that we need to find a δ such that

$$\delta+(\sqrt{1+\delta}-1)(\frac{L}{L_{\max}}+1) < \eta\gamma. \tag{B.27}$$

In other words, if we choose

$$\delta^* = \frac{\eta\gamma}{1+\frac{L}{L_{\max}}}$$

then we can conclude that $\mathbf{G}(\delta^*)$ has a negative eigenvalue which is less than $-\eta\gamma+\delta^* = -\frac{\eta\gamma}{1+\frac{L_{\max}}{L}}$.

In the following, we will check that δ^* is a valid choice, meaning that equation (B.27) holds when $\delta^* = \frac{\eta\gamma}{1+\frac{L}{L_{\max}}}$.

First step : since $L/L_{\max} \geq 1$, we have $\eta\gamma/(1+L/L_{\max}) \leq \eta\gamma/2$.

Second step : we only need to check

$$(\sqrt{1+\delta}-1)(\frac{L}{L_{\max}}+1) < \frac{\eta\gamma}{2},$$

meaning that it is sufficient to check

$$(\frac{L}{L_{\max}}+1)^2(1+\delta) \leq \left(\frac{L}{L_{\max}}+1+\frac{\eta\gamma}{2}\right)^2. \tag{B.28}$$

It can be easily check that the left-hand side (LHS) of (B.28) with chosen δ^* is

$$(\frac{L}{L_{\max}}+1)^2(1+\frac{\eta\gamma}{\frac{L}{L_{\max}}+1}) \leq (\frac{L}{L_{\max}}+1)^2+(\frac{L}{L_{\max}}+1)^2\eta\gamma < (\frac{L}{L_{\max}}+1)^2+(\frac{L}{L_{\max}}+1)^2\eta\gamma+\frac{\eta^2\gamma^2}{4},$$

which is RHS of (B.28).

Therefore, we can conclude that $\mathbf{Q}(1+\delta^*)$ has a negative eigenvalue.

When δ is large, it is easy to check $\mathbf{Q}(1+\delta)$ has a positive eigenvalue, since term $\delta^2\mathbf{I}$ dominates the spectrum of matrix $\mathbf{Q}(1+\delta)$ in (B.24). Since the eigenvalue is continuous with respect to δ, we can conclude there exists a largest δ, i.e., $\widehat{\delta}$, such that $\mathbf{Q}(1+\widehat{\delta})$ has a zero eigenvalue, i.e., $\det(\mathbf{Q}(1+\widehat{\delta}))=0$ where $1+\widehat{\delta}$ is at least

$$1+\delta^* = 1+\frac{\eta\gamma}{L/L_{\max}+1}. \tag{B.29}$$

Therefore, we can conclude that there exits a largest real eigenvalue of $\mathbf{M}^{-1}\mathbf{T}$ which is $1+\widehat{\delta} > 1+\delta^* > 1$. □

B.2.3 Proof of Lemma 5

Proof. Under Assumption 1, we have (descent lemma)

$$\begin{aligned} f(\mathbf{x}^{(t+1)}) \leq & f(\mathbf{x}^{(t)}) + \sum_{k=1}^{2} \nabla_k f(\mathbf{h}_{-k}^{(t)}, \mathbf{x}_k^{(t)})^T(\mathbf{x}_k^{(t+1)} - \mathbf{x}_k^{(t)}) + \sum_{k=1}^{2}\frac{L_k}{2}\|\mathbf{x}_k^{(t+1)} - \mathbf{x}_k^{(t)}\|^2 \\ \overset{(a)}{\leq} & f(\mathbf{x}^{(t)}) - \sum_{k=1}^{2}\eta\|\nabla_k f(\mathbf{h}_{-k}^{(t)}, \mathbf{x}_k^{(t)})\|^2 + \sum_{k=1}^{2}\frac{\eta^2 L_k}{2}\|\nabla_k f(\mathbf{h}_{-k}^{(t)}, \mathbf{x}_k^{(t)})\|^2 \\ \overset{(b)}{\leq} & f(\mathbf{x}^{(t)}) - \sum_{k=1}^{2}\frac{\eta}{2}\|\nabla_k f(\mathbf{h}_{-k}^{(t)}, \mathbf{x}_k^{(t)})\|^2 \end{aligned} \tag{B.30}$$

where (a) is true because of the update rule of gradient descent in each block and Assumption 1, in (b) we used $\eta \leq 1/L_{\max}$. □

B.2.4 Proof of Lemma 14

Proof. Without loss of generality, let $\mathbf{u}^{(0)}$ be the origin, i.e., $\mathbf{u}^{(0)} = 0$. According to the AGD update rules, we have

$$\mathbf{u}^{(t+1)} = \mathbf{u}^{(t)} - \eta \begin{bmatrix} \nabla_1 f(\mathbf{u}_1^{(t)}, \mathbf{u}_2^{(t)}) \\ \nabla_2 f(\mathbf{u}_1^{(t+1)}, \mathbf{u}_2^{(t)}) \end{bmatrix}. \tag{B.31}$$

Then, we use the mathematical induction to prove that

$$\|\mathbf{u}^{(t)}\| \leq 5\widehat{c}\mathcal{S}, \forall t < T. \tag{B.32}$$

When $t = 0$, we have $\mathbf{u}^{(0)} = 0$, so (B.32) is true.

Suppose (B.32) is true for the case where $\tau \leq t$. We will show that (B.32) is also true for the case where $\tau = t + 1$.

First, we need to show the upper bound of $\|\mathbf{u}^{(t+1)} - \mathbf{u}^{(t)}\|$. According to the Taylor expansion and ρ-Hessian Lipschitz continuity, we have

$$f(\mathbf{u}^{(t)}) \leq f(\mathbf{u}^{(0)}) + \nabla f(\mathbf{u}^{(0)})^T(\mathbf{u}^{(t)} - \mathbf{u}^{(0)}) + \frac{1}{2}(\mathbf{u}^{(0)} - \mathbf{u}^{(t)})^T \nabla^2 f(\mathbf{u}^{(0)})(\mathbf{u}^{(0)} - \mathbf{u}^{(t)}) + \frac{\rho}{6}\|\mathbf{u}^{(t)} - \mathbf{u}^{(0)}\|^3.$$

Comparing with the definition of $\widehat{f}_{\mathbf{u}^{(0)}}(\mathbf{u}^{(t)})$, we have

$$\begin{aligned} |f(\mathbf{u}^{(t)}) - \widehat{f}_{\mathbf{u}^{(0)}}(\mathbf{u}^{(t)})| &\overset{(B.11)}{\leq} \frac{1}{2}(\mathbf{u}^{(0)} - \mathbf{u}^{(t)})^T \left(\nabla^2 f(\mathbf{u}^{(0)}) - \mathbf{H}\right)(\mathbf{u}^{(0)} - \mathbf{u}^{(t)}) + \frac{\rho}{6}\|\mathbf{u}^{(t)} - \mathbf{u}^{(0)}\|^3 \\ &\overset{(a)}{\leq} \frac{\rho}{2}\|\mathbf{u}^{(0)} - \widetilde{\mathbf{x}}^{(t)}\|\|\mathbf{u}^{(t)} - \mathbf{u}^{(0)}\|^2 + \frac{\rho}{6}\|\mathbf{u}^{(t)} - \mathbf{u}^{(0)}\|^3 \end{aligned}$$

where in (a) we also used ρ-Hessian Lipschitz continuity.

According to the definition of T, we know that $f(\mathbf{u}^{(0)}) - \widehat{f}_{\mathbf{u}^{(0)}}(\mathbf{u}^{(t)}) \leq 3\mathcal{F}$ for all $t < T$, which implies that

$$\begin{aligned}
f(\mathbf{u}^{(0)}) - f(\mathbf{u}^{(t)}) \leq & |f(\mathbf{u}^{(0)}) - \widehat{f}_{\mathbf{u}^{(0)}}(\mathbf{u}^{(t)})| + |\widehat{f}_{\mathbf{u}^{(0)}}(\mathbf{u}^{(t)}) - f(\mathbf{u}^{(t)})| \\
\overset{(B.13)}{\leq} & 3\mathcal{F} + \frac{\rho}{2}\|\widetilde{\mathbf{x}}^{(t)} - \mathbf{u}^{(0)}\|\|\mathbf{u}^{(t)} - \mathbf{u}^{(0)}\|^2 + \frac{\rho}{6}\|\mathbf{u}^{(t)} - \mathbf{u}^{(0)}\|^3 \\
\leq & 3\mathcal{F} + \frac{\rho}{2}\frac{\eta L_{\max}\mathcal{S}}{\kappa \log(\frac{d\kappa}{\delta})\mathcal{P}_1}(5\widehat{c}\mathcal{S})^2 + \frac{\rho}{6}(5\widehat{c}\mathcal{S})^3 \qquad \text{(B.33)} \\
\leq & 3\mathcal{F} + ((5\widehat{c})^2/4 + (5\widehat{c})^3/6)\rho\mathcal{S}^3 \\
\overset{(B.9c)}{\leq} & 3\mathcal{F} + \eta L_{\max}(5\widehat{c})^3\mathcal{F}\mathcal{P}_2^{-1} \qquad \text{(B.34)} \\
\leq & 4\mathcal{F} \qquad \text{(B.35)}
\end{aligned}$$

where in (B.35) we used $c_{\max} = \mathcal{P}_2/(5\widehat{c})^3$ and $\eta \leq c_{\max}/L_{\max}$.

From (B.30), we also know that

$$f(\mathbf{u}^{(t+1)}) \leq f(\mathbf{u}^{(t)}) - \frac{\eta}{2}\left(\|\nabla_1 f(\mathbf{u}_1^{(t)}, \mathbf{u}_2^{(t)})\|^2 + \|\nabla_2 f(\mathbf{u}_1^{(t+1)}, \mathbf{u}_2^{(t)})\|^2\right), \quad \forall t < T. \tag{B.36}$$

For simplification of expression, we define

$$\mathbf{z}_{-1}^{(t)} := \mathbf{u}_2^{(t)} \quad \text{and} \quad \mathbf{z}_{-2}^{(t)} := \mathbf{u}_1^{(t+1)}, \quad \forall t < T. \tag{B.37}$$

Summing up (B.36) for $\tau = 0, \ldots, t$, we have

$$f(\mathbf{u}^{(t)}) \leq f(\mathbf{u}^{(0)}) - \sum_{\tau=0}^{t-1}\sum_{k=1}^{2}\frac{\eta}{2}\|\nabla_k f(\mathbf{z}_{-k}^{(\tau)}, \mathbf{u}_k^{(\tau)})\|^2, \quad \forall t < T. \tag{B.38}$$

Combining (B.35) and (B.38), we know that

$$\sum_{\tau=0}^{t-1}\sum_{k=1}^{2}\frac{\eta}{2}\|\nabla_k f(\mathbf{z}_{-k}^{(\tau)}, \mathbf{u}_k^{(\tau)})\|^2 \leq 4\mathcal{F}, \tag{B.39}$$

which implies

$$\max_{\tau}\sum_{k=1}^{2}\frac{\eta}{2}\|\nabla_k f(\mathbf{z}_{-k}^{(\tau)}, \mathbf{u}_k^{(\tau)})\|^2 \leq 4\mathcal{F}, \tau \leq t-1. \tag{B.40}$$

According to (B.31), we know

$$
\begin{aligned}
&\|\mathbf{u}^{(t+1)}-\mathbf{u}^{(t)}\|^2\\
=&\eta^2\sum_{k=1}^{2}\|\nabla_k f(\mathbf{z}_{-k}^{(t)},\mathbf{u}_k^{(t)})\|^2\\
=&2\eta^2\sum_{k=1}^{2}\|\nabla_k f(\mathbf{z}_{-k}^{(t)},\mathbf{u}_k^{(t)})-\nabla_k f(\mathbf{z}_{-k}^{(t-1)},\mathbf{u}_k^{(t-1)})\|^2+2\eta^2\sum_{k=1}^{2}\|\nabla_k f(\mathbf{z}_{-k}^{(t-1)},\mathbf{u}_k^{(t-1)})\|^2\\
=&2\eta^2\left(2\sum_{k=1}^{2}\|\nabla_k f(\mathbf{z}_{-k}^{(t)},\mathbf{u}_k^{(t)})-\nabla_k f(\mathbf{z}_{-k}^{(t-1)},\mathbf{u}_k^{(t)})\|^2+2\sum_{k=1}^{2}\|\nabla_k f(\mathbf{z}_{-k}^{(t-1)},\mathbf{u}_k^{(t)})-\nabla_k f(\mathbf{z}_{-k}^{(t-1)},\mathbf{u}_k^{(t-1)})\|^2\right)\\
&+2\eta^2\sum_{k=1}^{2}\|\nabla_k f(\mathbf{z}_{-k}^{(t-1)},\mathbf{u}_k^{(t-1)})\|^2\\
\overset{(a)}{\leq}&8\eta^2L_{\max}^2\|\mathbf{u}^{(t+1)}-\mathbf{u}^{(t)}\|^2+4\eta^2L_{\max}^2\|\mathbf{u}^{(t)}-\mathbf{u}^{(t-1)}\|^2+16\eta\mathcal{F}.
\end{aligned}
$$

where in (a) we used Lipschitz continuity, i.e., $\sum_{k=1}^{2}\|\nabla_k f(\mathbf{z}_{-k}^{(t)},\mathbf{u}_k^{(t)})-\nabla_k f(\mathbf{z}_{-k}^{(t-1)},\mathbf{u}_k^{(t)})\|^2\leq L_{\max}^2\|\mathbf{u}_1^{(t+1)}-\mathbf{u}_1^{(t)}\|^2+L_{\max}^2\|\mathbf{u}_2^{(t)}-\mathbf{u}_2^{(t-1)}\|^2$, and $\sum_{k=1}^{2}\|\nabla_k f(\mathbf{z}_{-k}^{(t-1)},\mathbf{u}_k^{(t)})-\nabla_k f(\mathbf{z}_{-k}^{(t-1)},\mathbf{u}_k^{(t-1)})\|^2\leq L_{\max}^2\|\mathbf{u}_1^{(t+1)}-\mathbf{u}_1^{(t)}\|$.

Then, we have

$$
\begin{aligned}
\|\mathbf{u}^{(t+1)}-\mathbf{u}^{(t)}\|^2\leq&\underbrace{\frac{4\eta^2L_{\max}^2}{(1-8\eta^2L_{\max}^2)}}_{:=\omega}\|\mathbf{u}^{(t)}-\mathbf{u}^{(t-1)}\|^2+\frac{16\eta\mathcal{F}}{(1-8\eta^2L_{\max}^2)}\\
=&\omega^t\|\mathbf{u}^{(1)}-\mathbf{u}^{(0)}\|^2+\sum_{\tau=0}^{t-1}\omega^\tau\frac{16\eta\mathcal{F}}{(1-8\eta^2L_{\max}^2)}\\
\overset{(a)}{\leq}&\frac{1-\omega^t}{1-\omega}\frac{16\eta\mathcal{F}}{(1-8\eta^2L_{\max}^2)}\leq\frac{1}{1-\omega}\frac{16\eta\mathcal{F}}{(1-8\eta^2L_{\max}^2)}<1.14*16\eta\mathcal{F}<18.2\eta\mathcal{F}
\end{aligned}
$$

where (a) is true because we have $\|\mathbf{u}^{(1)}-\mathbf{u}^{(0)}\|^2\leq 16\eta\mathcal{F}$ since $t<T$ and (B.40), and we used $\eta\leq c'_{\max}/L_{\max}$ where $c'_{\max}=1/10$ such that $\omega\approx 0.0435<1$.

Then, we can obtain

$$
\|\mathbf{u}^{(t+1)}-\mathbf{u}^{(t)}\|\leq 4.3\sqrt{\eta\mathcal{F}}\overset{\text{(B.9a)}}{\leq}\frac{4.3\eta\mathcal{G}}{\kappa}. \tag{B.41}
$$

Based on (B.41), we can get the upper bound of the sum of $\|\mathbf{u}^{(t+1)}-\mathbf{u}^{(t)}\|, \forall t<T$ as the following,

$$
\sum_{\tau=1}^{t+1}\|\mathbf{u}^{(\tau)}-\mathbf{u}^{(\tau-1)}\|\leq\sqrt{t\sum_{\tau=1}^{t+1}\|\mathbf{u}^{(\tau)}-\mathbf{u}^{(\tau-1)}\|^2}\overset{\text{(B.41)}}{\leq}T\cdot\frac{4.3\eta\mathcal{G}}{\kappa}\leq\widehat{c}\mathcal{T}\frac{4.3\eta\mathcal{G}}{\kappa}\overset{\text{(B.9b)}}{\leq}4.3\widehat{c}\mathcal{S}, \tag{B.42}
$$

which implies

$$\|\mathbf{u}^{(t+1)}\| \overset{(a)}{\leq} \sum_{\tau=1}^{t+1} \|\mathbf{u}^{(\tau)} - \mathbf{u}^{(\tau-1)}\| + \|\mathbf{u}^{(0)}\| \leq 4.3\widehat{c}\mathcal{S} \tag{B.43}$$

where in (a) we used the triangle inequality and $\mathbf{u}^{(0)} = 0$.

Due to the following fact

$$\|\mathbf{u}^{(t+1)} - \widetilde{\mathbf{x}}^{(t)}\| = \|\mathbf{u}^{(t+1)} - \mathbf{u}^{(0)} + \mathbf{u}^{(0)} - \widetilde{\mathbf{x}}^{(t)}\| \leq \|\mathbf{u}^{(t+1)} - \mathbf{u}^{(0)}\| + \|\mathbf{u}^{(0)} - \widetilde{\mathbf{x}}^{(t)}\| \leq 4.3\widehat{c}\mathcal{S} + \mathcal{S}/(2\kappa \log(\frac{d\kappa}{\delta})), \tag{B.44}$$

we have $\|\mathbf{u}^{(t+1)} - \widetilde{\mathbf{x}}^{(t)}\| \leq 5\widehat{c}\mathcal{S}$ since $\widehat{c} \geq 2$. Therefore, we know that there exits $c_{\max}^{(1)} = \min\{c_{\max}, c'_{\max}\}$ such that $\|\mathbf{u}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\| \leq 5\widehat{c}\mathcal{S}, \forall t < T$ when $\eta \leq c_{\max}^{(1)}/L_{\max}$, which completes the proof. □

B.2.5 Proof of Lemma 15

Proof. Let $\mathbf{u}^{(0)} = 0$ and define $\mathbf{v}^{(t)} := \mathbf{w}^{(t)} - \mathbf{u}^{(t)}$. According to the assumption of Lemma 15, we know that $\mathbf{v}^{(0)} = \upsilon[\eta L_{\max}\mathcal{S}/(\kappa \log(\frac{d\kappa}{\delta})\mathcal{P}_1)]\vec{\mathbf{e}}$ when $\upsilon \in [\delta/(2\sqrt{d}), 1]$. First, we define an auxiliary function

$$h(\theta) := \begin{bmatrix} \nabla_1 f(\mathbf{u}_1^{(t)} + \theta\mathbf{v}_1^{(t)}, \mathbf{u}_2^{(t)} + \theta\mathbf{v}_2^{(t)}) \\ \nabla_2 f(\mathbf{u}_1^{(t+1)} + \theta\mathbf{v}_1^{(t+1)}, \mathbf{u}_2^{(t)} + \theta\mathbf{v}_2^{(t)}) \end{bmatrix},$$

then have

$$h(0) = \begin{bmatrix} \nabla_1 f(\mathbf{u}_1^{(t)}, \mathbf{u}_2^{(t)}) \\ \nabla_2 f(\mathbf{u}_1^{(t+1)}, \mathbf{u}_2^{(t)}) \end{bmatrix}, \quad h(1) = \begin{bmatrix} \nabla_1 f(\mathbf{u}_1^{(t)} + \mathbf{v}_1^{(t)}, \mathbf{u}_2^{(t)} + \mathbf{v}_2^{(t)}) \\ \nabla_2 f(\mathbf{u}_1^{(t+1)} + \mathbf{v}_1^{(t+1)}, \mathbf{u}_2^{(t)} + \mathbf{v}_2^{(t)}) \end{bmatrix},$$

$$g(\theta) = \frac{dh(\theta)}{d\theta} = \underbrace{\begin{bmatrix} \nabla_{11}^2 f(\mathbf{u}_1^{(t)} + \theta\mathbf{v}_1^{(t)}, \mathbf{u}_2^{(t)} + \theta\mathbf{v}_2^{(t)}) & \nabla_{12}^2 f(\mathbf{u}_1^{(t)} + \theta\mathbf{v}_1^{(t)}, \mathbf{u}_2^{(t)} + \theta\mathbf{v}_2^{(t)}) \\ 0 & \nabla_{22}^2 f(\mathbf{u}_1^{(t+1)} + \theta\mathbf{v}_1^{(t+1)}, \mathbf{u}_2^{(t)} + \theta\mathbf{v}_2^{(t)}) \end{bmatrix}}_{\widetilde{\boldsymbol{\mathcal{H}}}_u^{(t)}(\theta)} \mathbf{v}^{(t)}$$

$$+ \underbrace{\begin{bmatrix} 0 & 0 \\ \nabla_{21}^2 f(\mathbf{u}_1^{(t+1)} + \theta\mathbf{v}_1^{(t+1)}, \mathbf{u}_2^{(t)} + \theta\mathbf{v}_2^{(t)}) & 0 \end{bmatrix}}_{\widetilde{\boldsymbol{\mathcal{H}}}_l^{(t)}(\theta)} \mathbf{v}^{(t+1)},$$

$$\begin{bmatrix} \nabla_1 f(\mathbf{w}_1^{(t)}, \mathbf{w}_2^{(t)}) \\ \nabla_2 f(\mathbf{w}_1^{(t+1)}, \mathbf{w}_2^{(t)}) \end{bmatrix} = \int_0^1 g(\theta) d\theta + \begin{bmatrix} \nabla_1 f(\mathbf{u}_1^{(t)}, \mathbf{u}_2^{(t)}) \\ \nabla_2 f(\mathbf{u}_1^{(t+1)}, \mathbf{u}_2^{(t)}) \end{bmatrix}.$$

Then, we consider sequence $\mathbf{w}^{(t)}$, i.e.,

$$\mathbf{u}^{(t+1)} + \mathbf{v}^{(t+1)} = \mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \begin{bmatrix} \nabla_1 f(\mathbf{w}_1^{(t)}, \mathbf{w}_2^{(t)}) \\ \nabla_2 f(\mathbf{w}_1^{(t+1)}, \mathbf{w}_2^{(t)}) \end{bmatrix} \tag{B.45}$$

$$= \mathbf{u}^{(t)} + \mathbf{v}^{(t)} - \eta \begin{bmatrix} \nabla_1 f(\mathbf{u}_1^{(t)} + \mathbf{v}_1^{(t)}, \mathbf{u}_2^{(t)} + \mathbf{v}_2^{(t)}) \\ \nabla_2 f(\mathbf{u}_1^{(t+1)} + \mathbf{v}_1^{(t+1)}, \mathbf{u}_2^{(t)} + \mathbf{v}_2^{(t)}) \end{bmatrix}$$

$$= \mathbf{u}^{(t)} + \mathbf{v}^{(t)} - \eta \begin{bmatrix} \nabla_1 f(\mathbf{u}_1^{(t)}, \mathbf{u}_2^{(t)}) \\ \nabla_2 f(\mathbf{u}_1^{(t+1)}, \mathbf{u}_2^{(t)}) \end{bmatrix} - \int_0^1 g(\theta) d\theta \tag{B.46}$$

$$\overset{(a)}{=} \mathbf{u}^{(t)} + \mathbf{v}^{(t)} - \eta \begin{bmatrix} \nabla_1 f(\mathbf{u}_1^{(t)}, \mathbf{u}_2^{(t)}) \\ \nabla_2 f(\mathbf{u}_1^{(t+1)}, \mathbf{u}_2^{(t)}) \end{bmatrix} - \eta\widetilde{\Delta}_u^{(t)}\mathbf{v}^{(t)} - \eta\mathbf{H}_u\mathbf{v}^{(t)} - \eta\widetilde{\Delta}_l^{(t)}\mathbf{v}^{(t+1)} - \eta\mathbf{H}_l\mathbf{v}^{(t+1)} \tag{B.47}$$

where in (a) we used the following definitions:

$$\widetilde{\Delta}_u^{(t)} := \int_0^1 \widetilde{\boldsymbol{\mathcal{H}}}_u^{(t)}(\theta) d\theta - \mathbf{H}_u, \tag{B.48}$$

$$\widetilde{\Delta}_l^{(t)} := \int_0^1 \widetilde{\boldsymbol{\mathcal{H}}}_l^{(t)}(\theta) d\theta - \mathbf{H}_l, \tag{B.49}$$

and

$$\mathbf{H}_u := \begin{bmatrix} \nabla^2_{11} f(\widetilde{\mathbf{x}}^{(t)}) & \nabla^2_{12} f(\widetilde{\mathbf{x}}^{(t)}) \\ 0 & \nabla^2_{22} f(\widetilde{\mathbf{x}}^{(t)}) \end{bmatrix} \quad \mathbf{H}_l := \begin{bmatrix} 0 & 0 \\ \nabla^2_{21} f(\widetilde{\mathbf{x}}^{(t)}) & 0 \end{bmatrix}. \tag{B.50}$$

Obviously, $\mathbf{H} = \mathbf{H}_l + \mathbf{H}_u$.

Dynamics of $\mathbf{v}^{(t)}$: Since the first two terms at RHS of (B.47) combined with $\mathbf{u}^{(t)}$ at LHS of (B.47) are exactly the same as (B.31). It can be observed that equation (B.47) gives the dynamic of $\mathbf{v}^{(t)}$, i.e.,

$$\mathbf{v}^{(t+1)} = \mathbf{v}^{(t)} - \eta\widetilde{\Delta}_u^{(t)}\mathbf{v}^{(t)} - \eta\mathbf{H}_u\mathbf{v}^{(t)} - \eta\widetilde{\Delta}_l^{(t)}\mathbf{v}^{(t+1)} - \eta\mathbf{H}_l\mathbf{v}^{(t+1)}. \tag{B.51}$$

Then, we can rewrite (B.51) in a matrix form as the following.

$$\underbrace{(\mathbf{I} + \eta\mathbf{H}_l)}_{:=\mathbf{M}}\mathbf{v}^{(t+1)} + \eta\widetilde{\Delta}_l^{(t)}\mathbf{v}^{(t+1)} \overset{\text{(B.47)}}{=} \underbrace{(\mathbf{I} - \eta\mathbf{H}_u)}_{:=\mathbf{T}}\mathbf{v}^{(t)} - \eta\widetilde{\Delta}_u^{(t)}\mathbf{v}^{(t)}. \tag{B.52}$$

It is worth noting that matrix $\mathbf{M}$ is a lower triangular matrix where the diagonal entries are all 1s, so it is invertible.

Taking the inverse of $\mathbf{M}$ on both sides of (B.52), we can obtain

$$\mathbf{v}^{(t+1)} + \mathbf{M}^{-1}\eta\widetilde{\Delta}_l^{(t)}\mathbf{v}^{(t+1)} = \mathbf{M}^{-1}\mathbf{T}\widehat{\mathbf{v}}^{(t)} - \mathbf{M}^{-1}\eta\widetilde{\Delta}_u^{(t)}\mathbf{v}^{(t)}. \tag{B.53}$$

Let $\mathbb{P}_{\texttt{left}}$ denote the projection operator that projects the vector onto the space spanned by the eigenvector of $\mathbf{M}^{-1}\mathbf{T}$ whose eigenvalue is maximum. Taking the projection on both sides of (B.53), we have

$$\mathbb{P}_{\texttt{left}}\widehat{\mathbf{v}}^{(t+1)} + \mathbb{P}_{\texttt{left}}\mathbf{M}^{-1}\eta\widetilde{\Delta}_l^{(t)}\mathbf{v}^{(t+1)} = \mathbb{P}_{\texttt{left}}(\mathbf{M}^{-1}\mathbf{T})\widehat{\mathbf{v}}^{(t)} - \mathbb{P}_{\texttt{left}}\mathbf{M}^{-1}\eta\widetilde{\Delta}_u^{(t)}\mathbf{v}^{(t)}. \tag{B.54}$$

From Lemma 4, we know that the maximum eigenvalue of $\mathbf{M}^{-1}\mathbf{T}$ is greater than 1.

Relationship of the Norm of $\mathbf{v}^{(t)}$ Projected in the Two Subspaces: Let $\phi^{(t)}$ denote the norm of $\mathbf{v}^{(t)}$ projected onto the space spanned by the eigenvector of $\mathbf{M}^{-1}\mathbf{T}$ whose maximum eigenvalue is $1 + \widehat{\delta}$ where $\widehat{\delta} \geq \eta\gamma/(1 + L/L_{\max})$ due to Lemma 4, and $\theta^{(t)}$ denote the norm of $\mathbf{v}^{(t)}$

projected onto the remaining space. From (B.54), we can have

$$\phi^{(t+1)} \overset{(a)}{\geq} (1+\widehat{\delta})\phi^{(t)} - \eta\|\mathbf{M}^{-1}\|\|\widetilde{\Delta}_l^{(t)}\|\|\widehat{\mathbf{v}}^{(t+1)}\| - \eta\|\mathbf{M}^{-1}\|\|\widetilde{\Delta}_u^{(t)}\|\|\mathbf{v}^{(t)}\|, \tag{B.55}$$

$$\theta^{(t+1)} \leq (1+\widehat{\delta})\theta^{(t)} + \eta\|\mathbf{M}^{-1}\|\|\widetilde{\Delta}_l^{(t)}\|\|\widehat{\mathbf{v}}^{(t+1)}\| + \eta\|\mathbf{M}^{-1}\|\|\widetilde{\Delta}_u^{(t)}\|\|\mathbf{v}^{(t)}\|. \tag{B.56}$$

where (a) is true because we applied the triangle inequality since η is sufficiently small. Also, since $\mathbf{M}^{-1} = \mathbf{I} - \eta\mathbf{H}_l$, we have

$$\begin{aligned}
\|\mathbf{M}^{-1}\| \leq & 1 + \eta\|\mathbf{H}_l\| \\
\overset{(a)}{=} & 1 + \|\eta\mathbf{H}\odot\mathbf{D} - \eta\mathbf{H}_d\| \\
\leq & 1 + \eta\|\mathbf{H}\odot\mathbf{D}\| + \eta\|\mathbf{H}_d\| \\
\overset{(b)}{\leq} & 1 + \eta(1+\frac{1}{\pi}+\frac{\log(d)}{\pi})\|\mathbf{H}\| + \eta\|\mathbf{H}_d\| \\
\overset{(c)}{\leq} & 1 + \eta\log(2d)\|\mathbf{H}\| + \eta\|\mathbf{H}_d\| \\
\overset{(d)}{\leq} & 1 + \eta L\log(2d) + \eta L_{\max} \\
\leq & 1 + \frac{L}{L_{\max}}\log(2d) + 1 < 2(1+\frac{L\log(2d)}{L_{\max}})
\end{aligned} \tag{B.57}$$

where in (a) $\odot$ denotes the Hadamard product and

$$\mathbf{H}_d := \begin{bmatrix} \nabla_{11}^2 f(\widetilde{\mathbf{x}}^{(t)}) & 0 \\ 0 & \nabla_{22}^2 f(\widetilde{\mathbf{x}}^{(t)}) \end{bmatrix} \quad \mathbf{D} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \cdots & 1 & 1 \end{bmatrix} \in \mathbb{R}^{d\times d}$$

and inequality (b) comes from the result on the spectral norm of the triangular truncation operator (please see [Theorem 1](109)). In particular, by defining

$$Y(\mathbf{D}) := \max\left\{\frac{\|\mathbf{H}\odot\mathbf{D}\|}{\|\mathbf{H}\|}, \mathbf{H}\neq 0\right\},$$

we have

$$\left|\frac{Y(\mathbf{D})}{\log(d)} - \frac{1}{\pi}\right| \leq (1+\frac{1}{\pi})\frac{1}{\log(d)}, \tag{B.58}$$

(c) is true for $d \geq 3$, in (d) we used the fact that $\|\mathbf{H}\| \leq L$ and $\|\mathbf{H}_d\| \leq L_{\max}$.

Since $\|\mathbf{w}^{(0)} - \widetilde{\mathbf{x}}^{(t)}\| \leq \|\mathbf{u}^{(0)} - \widetilde{\mathbf{x}}^{(t)}\| + \|\mathbf{v}^{(0)}\| \leq 2r$, we can apply Lemma 14. Then, we know $\|\mathbf{w}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\| \leq 5\widehat{c}\mathcal{S}, \forall t < T$. According to the assumptions of Lemma 15, we have $\|\mathbf{u}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\| \leq 5\widehat{c}\mathcal{S}$, and

$$\|\mathbf{v}^{(t)}\| = \|\mathbf{w}^{(t)} - \mathbf{u}^{(t)}\| \leq \|\mathbf{u}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\| + \|\mathbf{w}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\| \leq 10\widehat{c}\mathcal{S}. \tag{B.59}$$

From (B.41), we know that

$$\|\mathbf{w}^{(t+1)} - \mathbf{w}^{(t)}\| \leq \frac{4.3\eta\mathcal{G}}{\kappa} = \frac{4.3\eta^3 L_{\max}^3 \frac{\gamma}{\rho}}{\kappa^2 \log^3 \frac{d\kappa}{\delta} \mathcal{P}_1^3 \mathcal{P}_2} \leq \mathcal{S},$$

since $\mathcal{P}_1 \geq 2$ and we choose $\eta \leq c_{\max}/L_{\max}$ and $c_{\max} = 1/10$. Similarly, we also have $\|\mathbf{u}^{(t+1)} - \mathbf{u}^{(t)}\| \leq \mathcal{S}$.

According to Lipsichiz continuity, we have the following bounds of $\|\mathbf{v}^{(t+1)}\|$, $\|\widetilde{\Delta}_u^{(t)}\|$ and $\|\widetilde{\Delta}_l^{(t)}\|$.

1. Relation between $\|\mathbf{v}^{(t)}\|$ and $\|\mathbf{v}^{(t+1)}\|$: We also know that

$$\begin{aligned}
\|\mathbf{v}^{(t+1)}\|^2 =& \|\mathbf{w}^{(t+1)} - \mathbf{u}^{(t+1)}\|^2 \\
=& \left\| \mathbf{w}^{(t)} - \eta \begin{bmatrix} \nabla_1 f(\mathbf{w}_1^{(t)}, \mathbf{w}_2^{(t)}) \\ \nabla_2 f(\mathbf{w}_1^{(t+1)}, \mathbf{w}_2^{(t)}) \end{bmatrix} - \left(\mathbf{u}^{(t)} - \eta \begin{bmatrix} \nabla_1 f(\mathbf{u}_1^{(t)}, \mathbf{u}_2^{(t)}) \\ \nabla_2 f(\mathbf{u}_1^{(t+1)}, \mathbf{u}_2^{(t)}) \end{bmatrix} \right) \right\|^2 \\
\leq& 2\|\mathbf{v}^{(t)}\|^2 + 4\eta^2 \left\| \begin{bmatrix} \nabla_1 f(\mathbf{w}_1^{(t)}, \mathbf{w}_2^{(t)}) \\ \nabla_2 f(\mathbf{w}_1^{(t+1)}, \mathbf{w}_2^{(t)}) \end{bmatrix} - \begin{bmatrix} \nabla_1 f(\mathbf{u}_1^{(t)}, \mathbf{w}_2^{(t)}) \\ \nabla_2 f(\mathbf{u}_1^{(t+1)}, \mathbf{w}_2^{(t)}) \end{bmatrix} \right\|^2 \\
& + 4\eta^2 \left\| \begin{bmatrix} \nabla_1 f(\mathbf{u}_1^{(t)}, \mathbf{w}_2^{(t)}) \\ \nabla_2 f(\mathbf{u}_1^{(t+1)}, \mathbf{w}_2^{(t)}) \end{bmatrix} - \begin{bmatrix} \nabla_1 f(\mathbf{u}_1^{(t)}, \mathbf{u}_2^{(t)}) \\ \nabla_2 f(\mathbf{u}_1^{(t+1)}, \mathbf{u}_2^{(t)}) \end{bmatrix} \right\|^2 \\
\overset{(a)}{\leq}& 2\|\mathbf{v}^{(t)}\|^2 + 4\eta^2 L_{\max}^2 (\|\mathbf{v}_1^{(t+1)}\|^2 + \|\mathbf{v}_1^{(t)}\|^2) + 8\eta^2 L_{\max}^2 \|\mathbf{v}_2^{(t)}\|^2
\end{aligned} \tag{B.60}$$

where (a) is true due to Lipschitz continuity.

We can express (B.60) as

$$(1 - 4\eta^2 L_{\max}^2)\|\mathbf{v}^{(t+1)}\| \leq (2 + 8\eta^2 L_{\max}^2)\|\mathbf{v}^{(t)}\|^2$$

which implies

$$\|\mathbf{v}^{(t+1)}\| \leq \sqrt{\frac{2 + \frac{8}{100}}{1 - \frac{4}{100}}} \|\mathbf{v}^{(t)}\| < \sqrt{2.2}\|\mathbf{v}^{(t)}\| < 1.5\|\mathbf{v}^{(t)}\|, \tag{B.61}$$

where we choose $\eta \leq c_{\max}/L_{\max}$ and $c_{\max} = 1/10$.

2. Bounds of $\|\widetilde{\Delta}_u^{(t)}\|$ and $\|\widetilde{\Delta}_l^{(t)}\|$:

According to ρ-Hessian Lipschitz continuity and Lemma 12, we have the size of $\widetilde{\Delta}_u^{(t)}$ as the following.

$$\begin{aligned}
\|(\widetilde{\Delta}_u^{(t)})\| \leq & \int_0^1 \|\widetilde{\boldsymbol{\mathcal{H}}}_u^{(t)}(\theta) - \mathbf{H}_u\| d\theta \\
\overset{(B.2)}{\leq} & \int_0^1 \rho \left(\|\mathbf{u}^{(t)} + \theta \mathbf{v}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\| + \left\| \begin{bmatrix} \mathbf{u}_1^{(t+1)} + \theta \mathbf{v}_1^{(t+1)} \\ \mathbf{u}_2^{(t)} + \theta \mathbf{v}_2^{(t)} \end{bmatrix} - \widetilde{\mathbf{x}}^{(t)} \right\| \right) d\theta \qquad \text{(B.62)} \\
\overset{(a)}{\leq} & \int_0^1 \rho \left(2\|\mathbf{u}^{(t)} + \theta \mathbf{v}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\| + \|\mathbf{u}^{(t+1)} + \theta \mathbf{v}^{(t+1)} - \widetilde{\mathbf{x}}^{(t)}\| \right) d\theta \\
\leq & \rho(\|\mathbf{u}^{(t+1)} - \widetilde{\mathbf{x}}^{(t)}\| + 2\|\mathbf{u}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\|) + \rho \int_0^1 \theta(\|\mathbf{v}^{(t+1)}\| + \|\mathbf{v}^{(t)}\|) d\theta \\
\leq & \rho \left(\|\mathbf{u}^{(t+1)} - \mathbf{u}^{(t)}\| + \|\mathbf{u}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\| + 2\|\mathbf{u}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\| \right) + 0.5\|\mathbf{v}^{(t+1)}\| + 0.5\|\mathbf{v}^{(t)}\| \Big) \\
\overset{(B.61)}{\leq} & \rho \left(\|\mathbf{u}^{(t+1)} - \mathbf{u}^{(t)}\| + 3\|\mathbf{u}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\| + 1.25\|\mathbf{v}^{(t)}\| \right) \\
\leq & \rho(1 + 27.5\widehat{c})\mathcal{S}
\end{aligned}$$

where (a) is true because

$$\begin{aligned}
& \left\| \begin{bmatrix} \mathbf{u}_1^{(t+1)} + \theta \mathbf{v}_1^{(t+1)} \\ \mathbf{u}_2^{(t)} + \theta \mathbf{v}_2^{(t)} \end{bmatrix} - \widetilde{\mathbf{x}}^{(t)} \right\| \\
\leq & \left\| \mathbf{I}_1 \left(\mathbf{u}^{(t+1)} + \theta \mathbf{v}^{(t+1)} - \widetilde{\mathbf{x}}^{(t)} \right) \right\| + \left\| \mathbf{I}_2 \left(\mathbf{u}^{(t)} + \theta \mathbf{v}^{(t)} - \widetilde{\mathbf{x}}^{(t)} \right) \right\| \qquad \text{(B.63)} \\
\overset{(B.5)}{\leq} & \|\mathbf{u}^{(t+1)} + \theta \mathbf{v}^{(t+1)} - \widetilde{\mathbf{x}}^{(t)}\| + \|\mathbf{u}^{(t)} + \theta \mathbf{v}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\|. \qquad \text{(B.64)}
\end{aligned}$$

Applying Lemma 12, we can also get the upper bound of $\|\widetilde{\Delta}_l^{(t)}\|$, i.e.,

$$\begin{aligned}
\|(\widetilde{\Delta}_l^{(t)})\| \leq & \int_0^1 \|\widetilde{\boldsymbol{\mathcal{H}}}_l^{(t)}(\theta) - \mathbf{H}_l\| d\theta \\
\overset{(B.3)}{\leq} & \int_0^1 \rho \left\| \begin{bmatrix} \mathbf{u}_1^{(t+1)} + \theta \mathbf{v}_1^{(t+1)} \\ \mathbf{u}_2^{(t)} + \theta \mathbf{v}_2^{(t)} \end{bmatrix} - \widetilde{\mathbf{x}}^{(t)} \right\| d\theta \\
\leq & \int_0^1 \rho(\|\mathbf{u}^{(t)} + \theta \mathbf{v}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\| + \|\mathbf{u}^{(t+1)} + \theta \mathbf{v}^{(t+1)} - \widetilde{\mathbf{x}}^{(t)})\| d\theta \\
\leq & \rho(\|\mathbf{u}^{(t+1)} - \widetilde{\mathbf{x}}^{(t)}\| + \|\mathbf{u}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\|) + \rho \int_0^1 \theta(\|\mathbf{v}^{(t+1)}\| + \|\mathbf{v}^{(t)}\|) d\theta \\
\overset{(B.61)}{\leq} & \rho \left(\|\mathbf{u}^{(t+1)} - \mathbf{u}^{(t)}\| + 2\|\mathbf{u}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\| + 1.25\|\mathbf{v}^{(t)}\| \right) \\
\leq & \rho(1 + 22.5\widehat{c})\mathcal{S}.
\end{aligned} \tag{B.65}$$

With the upper bounds of $\|\mathbf{v}^{(t+1)}\|$, $\|\widetilde{\Delta}_u^{(t)}\|$, $\|\widetilde{\Delta}_l^{(t)}\|$ and relation between $\|\mathbf{v}^{(t+1)}\|$ and $\|\mathbf{v}^{(t)}\|$, we can further simply (B.55) and (B.56) as follows,

$$\begin{aligned}
\phi^{(t+1)} &\overset{(B.55)}{\geq} (1 + \widehat{\delta})\phi^{(t)} - \eta(1.5\|\widetilde{\Delta}_l^{(t)}\| + \|\widetilde{\Delta}_u^{(t)}\|)\|\mathbf{M}^{-1}\|\|\mathbf{v}^{(t)}\| \\
\theta^{(t+1)} &\overset{(B.56)}{\leq} (1 + \widehat{\delta})\theta^{(t)} + \eta(1.5\|\widetilde{\Delta}_l^{(t)}\| + \|\widetilde{\Delta}_u^{(t)}\|)\|\mathbf{M}^{-1}\|\|\mathbf{v}^{(t)}\|
\end{aligned}$$

and further we have

$$\begin{aligned}
\phi^{(t+1)} &\geq (1 + \widehat{\delta})\phi^{(t)} - \eta(1.5\|\widetilde{\Delta}_l^{(t)}\| + \|\widetilde{\Delta}_u^{(t)}\|)\|\mathbf{M}^{-1}\|\sqrt{(\phi^{(t)})^2 + (\theta^{(t)})^2}, \\
\theta^{(t+1)} &\leq (1 + \widehat{\delta})\theta^{(t)} + \eta(1.5\|\widetilde{\Delta}_l^{(t)}\| + \|\widetilde{\Delta}_u^{(t)}\|)\|\mathbf{M}^{-1}\|\sqrt{(\phi^{(t)})^2 + (\theta^{(t)})^2},
\end{aligned}$$

since $\|\mathbf{v}^{(t)}\| = \sqrt{(\phi^{(t)})^2 + (\theta^{(t)})^2}$.

Consequently, we can arrive at

$$\phi^{(t+1)} \geq (1 + \widehat{\delta})\phi^{(t)} - \mu\sqrt{(\phi^{(t)})^2 + (\theta^{(t)})^2}, \tag{B.66}$$

$$\theta^{(t+1)} \leq (1 + \widehat{\delta})\theta^{(t)} + \mu\sqrt{(\phi^{(t)})^2 + (\theta^{(t)})^2}, \tag{B.67}$$

where μ is the upper bound of $\eta(1.5\|\widetilde{\Delta}_l^{(t)}\| + \|\widetilde{\Delta}_u^{(t)}\|)\|\mathbf{M}^{-1}\|$ and can be obtained by

$$\mu := \eta\rho\mathcal{S}\mathcal{P}_2(2.5 + 62\widehat{c}). \tag{B.68}$$

Quantifying the Norm of $\mathbf{v}^{(t)}$ Projected at Different Subspaces: Then, we will use mathematical induction to prove

$$\theta^{(t)} \leq 4\mu t \phi^{(t)}. \tag{B.69}$$

It is true when $t = 0$ since $\|\theta^{(0)}\| \overset{(B.14)}{=} 0$.

Assuming that equation (B.69) is true at the tth iteration, we need to prove

$$\theta^{(t+1)} \leq 4\mu(t+1)\phi^{(t+1)}. \tag{B.70}$$

Applying (B.66) into RHS of (B.70), we have

$$4\mu(t+1)\phi^{(t+1)} \geq 4\mu(t+1)\left((1+\widehat{\delta})\phi^{(t)} - \mu\sqrt{(\phi^{(t)})^2 + (\theta^{(t)})^2}\right) \tag{B.71}$$

and substituting (B.67) into LHS of (B.70), we have

$$\theta^{(t+1)} \leq (1+\widehat{\delta})(4\mu t\phi^{(t)}) + \mu\sqrt{(\phi^{(t)})^2 + (\theta^{(t)})^2}. \tag{B.72}$$

Then, our goal is to prove RHS of (B.71) is greater than RHS of (B.72). After some manipulations, it is sufficient to show

$$(1 + 4\mu(t+1))\left(\sqrt{(\phi^{(t)})^2 + (\theta^{(t)})^2}\right) \leq 4\phi^{(t)}. \tag{B.73}$$

In the following, we will show that the above relation is true.

First step : We know that

$$4\mu(t+1) \leq 4\mu T \overset{(B.68)}{\leq} 4\eta\rho\mathcal{S}\mathcal{P}_2(2.5+62\widehat{c})\widehat{c}\mathcal{T} \overset{(B.9d)(B.68)}{\leq} \frac{4\widehat{c}\eta^2 L_{\max}^2(2.5+62\widehat{c})}{\kappa\log(\frac{d\kappa}{\delta})\mathcal{P}_1} \overset{(a)}{\leq} 1 \tag{B.74}$$

where (a) is true because $\mathcal{P}_1 \geq 2$ and we choose $c'_{\max} = 1/(2\widehat{c}(2.5+62\widehat{c}))$ and $\eta \leq c'_{\max}/L_{\max}$.

Second step : Also, we know that

$$4\phi^{(t)} \geq 2\sqrt{2(\phi^{(t)})^2} \overset{(B.69),(B.74)}{\geq} (1+4\mu(t+1))\sqrt{(\phi^{(t)})^2 + (\theta^{(t)})^2}. \tag{B.75}$$

With the above two steps, we have $\theta^{(t+1)} \leq 4\mu(t+1)\phi^{(t+1)}$, which completes the induction.

Recursion of $\phi^{(t)}$:Using (B.69), we have $\theta^{(t)} \overset{(B.69)}{\leq} 4\mu t\phi^{(t)} \overset{(B.74)}{\leq} \phi^{(t)}$, which implies

$$\begin{aligned}\phi^{(t+1)} &\overset{(B.66)}{\geq} (1+\widehat{\delta})\phi^{(t)} - \mu\sqrt{(\phi^{(t)})^2+(\theta^{(t)})^2} \\ &\overset{(a)}{\geq} (1+\frac{\gamma\eta}{1+L/L_{\max}})\phi^{(t)} - \mu\sqrt{(\phi^{(t)})^2+(\theta^{(t)})^2} \\ &\overset{(b)}{\geq} (1+\frac{1}{1+L/L_{\max}}\frac{\gamma\eta}{2})\phi^{(t)} \end{aligned} \tag{B.76}$$

where in (a) we used Lemma 4, and (b) is true because

$$\begin{aligned}\mu =& \eta\rho\mathcal{S}\mathcal{P}_2(2.5+62\widehat{c}) \\ \leq& \frac{\gamma\eta}{1+L/L_{\max}}\frac{\eta^2 L_{\max}^2(2.5+62\widehat{c})}{\log^2(\frac{d\kappa}{\delta})\mathcal{P}_1} \\ \overset{(a)}{\leq}& \frac{1}{1+L/L_{\max}}\frac{\gamma\eta}{2\sqrt{2}}\end{aligned}$$

where in (a) we choose $c''_{\max} = 1/(2\sqrt{2}(2.5+62\widehat{c}))$ and $\eta \leq c''_{\max}/L_{\max}$.

Quantifying Escaping Time: From (B.59), we have

$$\begin{aligned}10\mathcal{S}\widehat{c} \geq& \|\mathbf{v}^{(t)}\| \geq \phi^{(t)} \\ \overset{(B.76)}{\geq}& (1+\frac{\gamma\eta}{2(1+L/L_{\max})})^t\phi^{(0)} \\ \overset{(a)}{\geq}& (1+\frac{\gamma\eta}{2(1+L/L_{\max})})^t\frac{\delta}{2\sqrt{d}}\frac{\eta L_{\max}\mathcal{S}}{\kappa}\log^{-1}(\frac{d\kappa}{\delta})\mathcal{P}_1^{-1} \\ \overset{(b)}{\geq}& (1+\frac{\gamma\eta}{2(1+L/L_{\max})})^t\frac{\delta}{2\sqrt{d}}\frac{c\mathcal{S}}{\kappa}\log^{-1}(\frac{d\kappa}{\delta})\mathcal{P}_1^{-1} \quad \forall t<T\end{aligned} \tag{B.77}$$

where in (a) we use condition $\upsilon \in [\delta/(2\sqrt{d}), 1]$, in (b) we used $\eta = c/L_{\max}$.

Since (B.77) is true for all $t < T$, we can have

$$\begin{aligned}T-1 \leq& \frac{\log(20\frac{\widehat{c}}{c}(\frac{\kappa\sqrt{d}}{\delta})\log(\frac{d\kappa}{\delta})\mathcal{P}_1)}{\log(1+\frac{\eta\gamma}{2(1+L/L_{\max})})} \\ \overset{(a)}{<}& \frac{4(1+L/L_{\max})\log(20(\frac{\sqrt{d}\kappa}{\delta})\frac{\widehat{c}}{c}\log(\frac{d\kappa}{\delta})\mathcal{P}_1)}{\eta\gamma} \\ \overset{(b)}{<}& \frac{4(1+L/L_{\max})\log(20(\frac{d\kappa}{\delta})^2\frac{\widehat{c}}{c}\mathcal{P}_1)}{\eta\gamma} \\ \overset{(c)}{<}& 4(2+\log(20\frac{\widehat{c}}{c}))\mathcal{T}\end{aligned} \tag{B.78}$$

where (a) comes from inequality $\log(1+x) > x/2$ when $x < 1$, in (b) we used relation $\log(x) < x, x > 0$, and (c) is true because $\delta \in (0, \frac{d\kappa}{e}]$ and $\log(d\kappa/\delta) > 1$ and $\mathcal{P}_1 > 1$ we have

$$\log(\frac{d\kappa}{\delta}\mathcal{P}_1) \leq \log(\frac{d\kappa}{\delta}) + \log(1 + \frac{L}{L_{\max}}) \leq \log(\frac{d\kappa}{\delta}) + \frac{L}{L_{\max}} \leq \log(\frac{d\kappa}{\delta})\mathcal{P}_1.$$

From (B.78), we know that

$$T < 4(2 + \log(20\frac{\widehat{c}}{c}))\mathcal{T} + 1 \overset{(a)}{<} 4(2\frac{1}{4} + \log(20\frac{\widehat{c}}{c})\mathcal{T} \tag{B.79}$$

where (a) is true due to the fact that $\eta L_{\max} \geq 1$, $\log(d\kappa/\delta) > 1$ and $\mathcal{P}_1 > 1$ so we know $\mathcal{T} \geq 1$.

When

$$4(2.25 + \log(20\frac{\widehat{c}}{c})) \leq \widehat{c}, \tag{B.80}$$

we will have $T < \widehat{c}\mathcal{T}$ where $c_{\max}^{(2)} := \min\{c_{\max}, c'_{\max}, c''_{\max}\}$.

Since $\widehat{c} \geq 2$, we have $c_{\max} = \min\{c_{\max}^{(1)}, c_{\max}^{(2)}\} \leq 1/(5\widehat{c})^3$. Also, we know that $c \leq c_{\max}$. Combining with (B.80), we need

$$\frac{\widehat{c}}{2^{\frac{\widehat{c}}{4} - 2.25 - \log(20)}} \leq c \leq \frac{1}{(5\widehat{c})^3}, \tag{B.81}$$

meaning that

$$125(2^{2.25+\log(20)}\widehat{c}^4) \leq 2^{\frac{\widehat{c}}{4}}. \tag{B.82}$$

It can be observed that LHS of (B.82) is a polynomial with respect to $\widehat{c}$ and RHS of (B.82) is a exponential function in terms of $\widehat{c}$, implying there exists a universal $\widehat{c}$ such that (B.82) holds. The proof is complete. □

B.2.6 Proof of Lemma 16

Proof. The proof of Lemma 16 is similar as the one of proving convergence of PGD shown in (101, Lemma 14,15). Considering the completeness of the whole proof in the dissertation, here we give the following proof of this lemma in details.

First, after the random perturbation, the objective function value in the worst case is increased at most by

$$\begin{aligned} f(\mathbf{u}^{(0)}) - f(\widetilde{\mathbf{x}}^{(t)}) \leq & \sum_{k=1}^{2} \nabla_k f(\widetilde{\mathbf{h}}_{-k}^{(t)}, \widetilde{\mathbf{x}}_k^{(t)})^T \xi_k + \frac{L_k}{2}\|\xi_k\|^2 \\ \leq & \sum_{k=1}^{2} \|\nabla_k f(\widetilde{\mathbf{h}}_{-k}^{(t)}, \widetilde{\mathbf{x}}_k^{(t)})\|\|\xi_k\| + \frac{L_{\max}}{2}\|\xi\|^2 \\ \overset{(a)}{\leq} & \|\xi\| \sqrt{\sum_{k=1}^{2} 2\|\nabla_k f(\widetilde{\mathbf{h}}_{-k}^{(t)}, \widetilde{\mathbf{x}}_k^{(t)})\|^2} + \frac{L_{\max}}{2}\|\xi\|^2 \\ \overset{(b)}{\leq} & \frac{\mathcal{G}}{\kappa}\frac{\eta L_{\max}\mathcal{S}}{\kappa \log(\frac{d\kappa}{\delta})\mathcal{P}_1} + \frac{L_{\max}}{2}(\frac{\eta L_{\max}\mathcal{S}}{\kappa \log(\frac{d\kappa}{\delta})\mathcal{P}_1})^2 \leq \frac{3}{2}\mathcal{F} \end{aligned} \tag{B.83}$$

where $\mathbf{u}^{(0)}$ is a vector that follows uniform distribution within the ball $\mathbb{B}^{(d)}_{\widetilde{\mathbf{x}}^{(t)}}(r)$, $\mathbb{B}^{(d)}_{\widetilde{\mathbf{x}}^{(t)}}$ denotes the d-dimensional ball centered at $\widetilde{\mathbf{x}}^{(t)}$ with radius r, ξ_k represents the kth block of the vector which is the difference between random generated vector $\mathbf{u}^{(0)}$ and $\widetilde{\mathbf{x}}^{(t)}$, and (a) is true because $\xi := [\xi_1, \ldots, \xi_K]$, $\|\xi_k\| \leq \|\xi\|, \forall k$, and in (b) we used $\kappa > 1$, $\log(d\kappa/\delta) > 1$ and Condition 1.

Second, under Assumption 1, let $\widetilde{\mathbf{x}}^{(t)}$ satisfy conditions Condition 1, and two PA-GD iterates $\{\mathbf{u}^{(t)}\}$ $\{\mathbf{w}^{(t)}\}$ satisfy the conditions as in Lemma 15. Selecting $c_{\max} = \min\{c_{\max}^{(1)}, c_{\max}^{(2)}\}$, so we have that $\eta \leq c_{\max}/L_{\max}$ is small enough such that Lemma 14 and Lemma 15 can both hold.

Let $T^* := \widehat{c}\mathcal{T}$ and $T' := \inf_t\{t | \widehat{f}_{\mathbf{u}^{(0)}}(\mathbf{u}^{(t)}) - f(\mathbf{u}^{(0)}) \leq -3\mathcal{F}\}$. Then, we have the following two cases to analyze the decrease of the objective value after T iterations with the random perturbation.

1. Case $T' \leq T^*$:

$$\begin{aligned} f(\mathbf{u}^{(T')}) - f(\mathbf{u}^{(0)}) \leq & \nabla f(\mathbf{u}^{(0)})^T(\mathbf{u}^{(T')} - \mathbf{u}^{(0)}) + \frac{1}{2}(\mathbf{u}^{(T')} - \mathbf{u}^{(0)})^T\nabla^2 f(\mathbf{u}^{(0)})(\mathbf{u}^{(T')} - \mathbf{u}^{(0)}) \\ & + \frac{\rho}{6}\|\mathbf{u}^{(T')} - \mathbf{u}^{(0)}\|^3 \\ \leq & \widehat{f}_{\mathbf{u}^{(0)}}(\mathbf{u}^{(t)}) - f(\mathbf{u}^{(0)}) + \frac{\rho}{2}\|\mathbf{u}^{(0)} - \widetilde{\mathbf{x}}^{(t)}\|\|\mathbf{u}^{(T')} - \mathbf{u}^{(0)}\|^2 \\ & + \frac{\rho}{6}\|\mathbf{u}^{(T')} - \mathbf{u}^{(0)}\|^3 \\ \overset{(B.33)-(B.34)}{\leq} & -3\mathcal{F} + 0.5\rho\mathcal{S}^3 \overset{(B.9c)}{\leq} -2.5\mathcal{F}. \end{aligned} \tag{B.84}$$

Based on Lemma 5, we know that AGD is always decreasing the objective function. For any $T \geq \mathcal{T}/c_{\max} \geq \widehat{c}\mathcal{T} = T^* \geq T'$, we have

$$f(\mathbf{u}^{(T)}) - f(\mathbf{u}^{(0)}) \leq f(\mathbf{u}^{(T^*)}) - f(\mathbf{u}^{(0)}) \leq f(\mathbf{u}^{(T')}) - f(\mathbf{u}^{(0)}) \leq -2.5\mathcal{F}$$

where $c_{\max} = \min\{1, 1/\widehat{c}\}$.

2. Case $T' > T^*$: Applying Lemma 14, we know that $\|\mathbf{u}^{(t)} - \mathbf{u}^{(0)}\| \leq 5\widehat{c}\mathcal{S}$ for $t \leq T^*$. Define $T'' = \inf_t\{t | \widehat{f}_{\mathbf{w}^{(0)}}(\mathbf{w}^{(t)}) - f(\mathbf{w}^{(0)}) \leq -3\mathcal{F}\}$. Then, after applying Lemma 15, we know $T'' \leq T^*$. Similar as (B.84), for $T \geq 1/c_{\max}\mathcal{T}$, we also have $f(\mathbf{w}^{(T)}) - f(\mathbf{w}^{(0)}) \leq f(\mathbf{w}^{T^*}) - f(\mathbf{w}^{(0)}) \leq f(\mathbf{w}^{T''}) - f(\mathbf{w}^{(0)}) \leq -2.5\mathcal{F}$.

Combining the above two cases, we have

$$\min\{f(\mathbf{u}^{(T)}) - f(\mathbf{u}^{(0)}), f(\mathbf{w}^{(T)}) - f(\mathbf{w}^{(0)})\} \leq -2.5\mathcal{F}, \tag{B.85}$$

meaning that at least one of the sequences can give a sufficient decrease of the objective function if the initial points of the two sequences are separated apart with each other far enough along direction $\vec{\mathbf{e}}$.

Therefore, we can conclude that if $\mathbf{u}^{(0)} \in \mathcal{X}_{\mathsf{stuck}}$, then $(\mathbf{u}^{(0)} \pm \upsilon r\vec{\mathbf{e}}) \notin \mathcal{X}_{\mathsf{stuck}}$ where $\upsilon \in [\frac{\delta}{2\sqrt{d}}, 1]$.

Finally, we give the upper bound of the volume of $\mathcal{X}_{\mathsf{stuck}}$,

$$\begin{aligned}
\mathsf{Vol}(\mathcal{X}_{\mathsf{stuck}}) &= \int_{\mathbb{B}^{(d)}_{\widetilde{\mathbf{x}}^{(t)}}} d\mathbf{u} I_{\mathcal{X}_{\mathsf{stuck}}}(\mathbf{u}) = \int_{\mathbb{B}^{(d-1)}_{\widetilde{\mathbf{x}}^{(t)}}} du_{-1} \int_{\widetilde{x}_1^{(t)} - \sqrt{r^2 - \|\widetilde{\mathbf{x}}_{-1}^{(t)} - u_{-1}\|^2}}^{\widetilde{x}_1^{(t)} + \sqrt{r^2 - \|\widetilde{\mathbf{x}}_{-1}^{(t)} - u_{-1}\|^2}} du_1 I_{\mathcal{X}_{\mathsf{stuck}}}(\mathbf{u}) \\
&\leq \int_{\mathbb{B}^{(d-1)}_{\widetilde{\mathbf{x}}^{(t)}}} du_{-1} \left(2\frac{\delta}{2\sqrt{d}r}\right) = \mathsf{Vol}(\mathbb{B}^{(d-1)}_{\widetilde{\mathbf{x}}^{(t)}}(r))\frac{r\delta}{\sqrt{d}}
\end{aligned}$$

where $I_{\mathsf{stuck}}(\mathbf{u})$ is an indicator function showing that $\mathbf{u}$ belongs to set $\mathcal{X}_{\mathsf{stuck}}$, and u_1 represents the component of vector $\mathbf{u}$ along $\vec{\mathbf{e}}$ direction, and $\mathbf{u}_{-1}$ is the remaining $d-1$ dimensional vector.

Then, the ratio of $\mathsf{Vol}(\mathcal{X}_{\mathsf{stuck}})$ over the whole volume of the perturbation ball can be upper bounded by

$$\frac{\mathsf{Vol}(\mathcal{X}_{\mathsf{stuck}})}{\mathsf{Vol}(\mathbb{B}^{(d)}_{\widetilde{\mathbf{x}}^{(t)}}(r))} \leq \frac{\frac{r\delta}{\sqrt{d}}\mathsf{Vol}(\mathbb{B}^{(d-1)}_{\widetilde{\mathbf{x}}^{(t)}}(r))}{\mathsf{Vol}(\mathbb{B}^{(d)}_{\widetilde{\mathbf{x}}^{(t)}}(r))} = \frac{\delta}{\sqrt{d\pi}}\frac{\Gamma(\frac{d}{2}+1)}{\Gamma(\frac{d}{2}+1)} \leq \frac{\delta}{\sqrt{d\pi}}\sqrt{\frac{d}{2}+\frac{1}{2}} \leq \delta$$

where $\Gamma(\cdot)$ denotes the Gamma function, and inequality is true due to the fact that $\Gamma(x+1)/\Gamma(x+1/2) < \sqrt{x+1/2}$ when $x \geq 0$.

Combining (B.83) and (B.85), we can show that

$$f(\mathbf{x}^{(T)}) - f(\widetilde{\mathbf{x}}^{(t)}) = f(\mathbf{x}^{(T)}) - f(\mathbf{u}^{(0)}) + f(\mathbf{u}^{(0)}) - f(\widetilde{\mathbf{x}}^{(t)}) \leq -2.5\mathcal{F} + 1.5\mathcal{F} \leq -\mathcal{F} \tag{B.86}$$

with at least probability $1 - \delta$. □

B.3 Proof of the Convergence Rate of PA-PP

First, we need to introduce some constants defined as follows,

$$\mathcal{F} := \eta^5 L_{\max}^5 \frac{\gamma^3}{\kappa^3 \rho^2} \log^{-6}\left(\frac{d\kappa}{\delta}\right) \mathcal{P}^{-2}, \qquad \mathcal{G} := \eta^2 L_{\max}^2 \frac{\gamma^2}{\rho} \log^{-3}\left(\frac{d\kappa}{\delta}\right) \mathcal{P}^{-1},$$
$$\mathcal{S} := \eta^2 L_{\max}^2 \frac{\gamma}{\kappa\rho} \log^{-2}\left(\frac{d\kappa}{\delta}\right) \mathcal{P}^{-1}, \qquad \mathcal{T} := \frac{\log\left(\frac{d\kappa}{\delta}\right)}{\eta\gamma}$$

where $\eta = 1/\nu$. In order to keep the completeness of the proof, the certain relations of these quantities are listed as follows, which are useful of simplifying the expressions in the proofs.

$$\sqrt{\mathcal{F}} = \frac{\sqrt{\eta}\mathcal{G}}{\kappa}, \tag{B.87a}$$
$$\frac{\eta\mathcal{G}\mathcal{T}}{\kappa} = \mathcal{S}, \tag{B.87b}$$
$$\rho\mathcal{S}^3 = \frac{\eta L_{\max}\mathcal{F}}{\mathcal{P}}, \tag{B.87c}$$
$$\eta\rho\mathcal{S}\mathcal{T} = \frac{\eta^2 L_{\max}^2}{\kappa \log(\frac{d\kappa}{\delta})\mathcal{P}}, \tag{B.87d}$$
$$\eta\rho\mathcal{S} = \eta L_{\max} \frac{\eta^2\gamma^2}{\log^2(\frac{d\kappa}{\delta})\mathcal{P}}. \tag{B.87e}$$

We also consider saddle point $\widetilde{\mathbf{x}}^{(t)}$ that satisfies the following condition.

Condition 2. *An ϵ-second order stationary point $\widetilde{\mathbf{x}}^{(t)}$ satisfies the following conditions:*

$$\|\mathbf{x}^{(t+1)} - \mathbf{x}^{(t)}\| \leq g_{th}/\nu \quad \textit{and} \quad \lambda_{\min}(\nabla^2 f(\widetilde{\mathbf{x}}^{(t)})) \leq -\gamma \tag{B.88}$$

where $g_{th} = \frac{\mathcal{G}}{2\kappa}$.

Then, we have the following preliminary lemmas.

Lemma 17. *If function $f(\cdot)$ is L-smooth with Lipschitz constant, the we have*

$$\|\nabla f(\mathbf{x}^{(t)})\|^2 \leq 4\nu\|\mathbf{x}^{(t+1)} - \mathbf{x}^{(t)}\|^2 \tag{B.89}$$

where sequence $\mathbf{x}_k^{(t)}, k = 1, 2$ is generated by the APP algorithm.

Lemma 18. *Under Assumption 1, we have block-wise Lipschitz continuity as the follows:*

$$\left\| \begin{bmatrix} \nabla^2_{11} f(\mathbf{x}) & \mathbf{0} \\ \nabla^2_{21} f(\mathbf{y}) & \nabla^2_{22} f(\mathbf{y}) \end{bmatrix} - \begin{bmatrix} \nabla^2_{11} f(\mathbf{z}) & \mathbf{0} \\ \nabla^2_{21} f(\mathbf{z}) & \nabla^2_{22} f(\mathbf{z}) \end{bmatrix} \right\| \leq \rho \left(\|\mathbf{x}-\mathbf{z}\| + \|\mathbf{y}-\mathbf{z}\| \right), \forall \mathbf{x}, \mathbf{y}, \mathbf{z} \tag{B.90}$$

and

$$\left\| \begin{bmatrix} \mathbf{0} & \nabla^2_{21} f(\mathbf{x}) \\ \mathbf{0} & \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{0} & \nabla^2_{12} f(\mathbf{y}) \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \right\| \leq \rho \|\mathbf{x}-\mathbf{y}\|, \forall \mathbf{x}, \mathbf{y}. \tag{B.91}$$

Second, we can have the descent lemma as the following

Lemma 19. *Under Assumption 1, for the APP algorithm with penalizer $\nu \geq 3L_{\max}$, we have*

$$f(\mathbf{x}^{(t+1)}) \leq f(\mathbf{x}^{(t)}) - \frac{\nu}{2} \|\mathbf{x}^{(t+1)} - \mathbf{x}^{(t)}\|^2.$$

Third, we need to characterize the convergence behaviour of PA-PP when $\|\mathbf{x}^{(t+1)} - \mathbf{x}^{(t)}\|$ is small. In this case, we need three steps to arrive the final results.

Step 1 : Quantify upper bound of the distance between generic iterate $\mathbf{u}^{(t)}$ and saddle point $\widetilde{\mathbf{x}}^{(t)}$.

Lemma 20. *Under Assumption 1, consider saddle point $\widetilde{\mathbf{x}}^{(t)}$ that satisfies Condition 2. For any constant $\widehat{c} \geq 2$, $\delta \in (0, \frac{d_\kappa}{e}]$, when initial point $\mathbf{u}^{(0)}$ satisfies*

$$\|\mathbf{u}^{(0)} - \widetilde{\mathbf{x}}^{(t)}\| \leq 2r, \tag{B.92}$$

then, with the definition of

$$r := \frac{\frac{L_{\max}}{\nu}\mathcal{S}}{\kappa \log(\frac{d_\kappa}{\delta})\mathcal{P}_1} \quad \text{and} \quad T := \min\{\inf_t \{t | \widehat{f}_{\mathbf{u}^{(0)}}(\mathbf{u}^{(t)}) - f(\mathbf{u}^{(0)}) \leq -3\mathcal{F}\}, \widehat{c}\mathcal{T}\}, \tag{B.93}$$

there exits constants $c^{(1)}_{\max}, \widehat{c}$ such that for any $\nu \geq L_{\max}/c^{(1)}_{\max}$, the iterates generated by PA-PP satisfy $\|\mathbf{u}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\| \leq 5\widehat{c}\mathcal{S}, \forall t < T$.

Step 2 : Quantify the escaping time of iterates near a strict saddle point.

Lemma 21. *Under Assumption 1, consider saddle point $\widetilde{\mathbf{x}}^{(t)}$ that satisfies satisfies Condition 2. There exist constants $c_{\max}^{(2)}$, $\widehat{c}$ such that: for any $\delta \in (0, \frac{d\kappa}{e}]$ and $\nu \geq L_{\max}/c_{\max}^{(2)}$, with the definition of*

$$T := \min\left\{\inf_t\{t|\widehat{f}_{\mathbf{w}_0}(\mathbf{w}^{(t)}) - f(\mathbf{w}^{(0)}) \leq -3\mathcal{F}\}, \widehat{c}\mathcal{T}\right\} \tag{B.94}$$

where two iterates $\{\mathbf{u}^{(t)}\}$ and $\{\mathbf{w}^{(t)}\}$ that are generated by PA-PP with initial points $\{\mathbf{u}^{(0)}, \mathbf{w}^{(0)}\}$ satisfying

$$\|\mathbf{u}^{(0)} - \widetilde{\mathbf{x}}^{(t)}\| \leq r, \quad \mathbf{w}^{(0)} = \mathbf{u}^{(0)} + \upsilon r\vec{\mathbf{e}}', \quad \upsilon \in [\delta/(2\sqrt{d}), 1], \tag{B.95}$$

where $\vec{\mathbf{e}}'$ denotes the eigenvector of $\mathbf{T}'^{-1}\mathbf{M}'$ whose corresponding positive eigenvalue is minimum, if $\|\mathbf{u}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\| \leq 5\widehat{c}\mathcal{S}, \forall t < T$, we will have $T < \widehat{c}\mathcal{T}$.

Step 3 : Quantify sufficient decrease with random perturbation. With Lemma 20 and Lemma 21, we can apply Lemma 16 directly and obtain the following lemma.

Lemma 22. *Under Assumption 1, there exists a universal constant $c_{\max}$, for any $\delta \in (0, d\kappa/e]$: consider a saddle point $\widetilde{\mathbf{x}}^{(t)}$ which satisfies (4.2), let $\mathbf{x}^{(0)} = \widetilde{\mathbf{x}}^{(t)} + \xi$ where ξ is generated randomly which follows the uniform distribution over a ball with radius r, and let $\mathbf{x}^{(t)}$ be the iterates of PA-PP starting from $\mathbf{x}^{(0)}$. Then, when step size $\nu \geq L_{\max}/c_{\max}$, with at least probability $1-\delta$, we have the following for any $T \geq \mathcal{T}/c_{\max}$*

$$f(\mathbf{x}^{(T)}) - f(\widetilde{\mathbf{x}}^{(t)}) \leq -\mathcal{F}. \tag{B.96}$$

Substituting $\nu = \frac{L_{\max}}{c}$,$\gamma = (L_{\max}\rho\epsilon)^{1/3}$, and $\delta = \frac{dL_{\max}}{(L_{\max}\rho\epsilon)^{1/3}}e^{-\chi}$ in to Lemma 22, we can obtain the following lemma immediately.

Lemma 23. *Under Assumption 1, there exists a absolute constant $c_{\max}$. Let $c \leq c_{\max}$, $\chi \geq 1$, and η, r, g_{th}, t_{th} calculated as Algorithm 3 describes. Let $\widetilde{\mathbf{x}}^{(t)}$ be a strict saddle point, which satisfies*

$$\|\nabla f(\widetilde{\mathbf{x}}^{(t)})\|^2 \leq 4\nu\|\mathbf{x}^{(t+1)} - \mathbf{x}^{(t)}\|^2 \leq 4g_{th}^2 \tag{B.97}$$

and

$$\lambda_{\min}(\nabla^2 f(\widetilde{\mathbf{x}}^{(t)})) \leq -\gamma.$$

Let $\mathbf{x}^{(t)} = \widetilde{\mathbf{x}}^{(t)} + \xi^{(t)}$ *where* $\xi^{(t)}$ *is generated randomly which follows the uniform distribution over* $\mathbb{B}_{\widetilde{\mathbf{x}}^{(t)}}(r)$*, and let* $\mathbf{x}^{(t+t_{th})}$ *be the iterates of PA-PP. With at least probability* $1 - \frac{dL_{\max}}{(L_{\max}\rho\epsilon)^{1/3}}e^{-\chi}$*, we have*

$$f(\mathbf{x}^{(t+t_{th})}) - f(\widetilde{\mathbf{x}}^{(t)}) \leq -f_{th}. \tag{B.98}$$

Finally, we can get the convergence rate of PA-PP as the following.

B.3.1 Proof of Corollary 3

Next, we prove the main theorem.

Proof. Submitting $\nu = \frac{L_{\max}}{c}$,$\gamma = (L_{\max}\rho\epsilon)^{1/3}$, and $\delta = \frac{dL_{\max}}{(L_{\max}\rho\epsilon)^{1/3}}e^{-\chi}$ into the definition of $\mathcal{F}, \mathcal{G}, \mathcal{T}$, we will have the following definitions.

$$\begin{aligned} f_{\mathsf{th}} :=& \mathcal{F} = \frac{c^5\epsilon^2}{L_{\max}\chi^6\mathcal{P}^2}, \\ g_{\mathsf{th}} :=& \frac{\mathcal{G}}{2\kappa} = \frac{c^2\epsilon}{2\chi\mathcal{P}}, \\ t_{\mathsf{th}} :=& \frac{\mathcal{T}}{c} = \frac{L_{\max}\chi}{c^2(L_{\max}\rho\epsilon)^{\frac{1}{3}}}. \end{aligned} \tag{B.99}$$

After applying Lemma 13, we know that

$$\|\nabla f(\mathbf{x})\| \leq \frac{c}{\chi^3\mathcal{P}}\epsilon \tag{B.100}$$

where $c \leq 1, \chi, \mathcal{P} \geq 1$.

Similarly, at any iteration, we need to consider two cases (we use the first iteration as an example):

1. In this case the gradient is large such that $\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\| > g_{\mathsf{th}}/\nu$: According to Lemma 19, we have

$$\begin{aligned} f(\mathbf{x}^{(1)}) - f(\mathbf{x}^{(0)}) &\leq -\frac{\nu}{2}\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\|^2 \leq -\frac{\nu}{2}g_{\mathsf{th}}^2 \\ &\overset{(a)}{=} -\frac{c^5}{8\chi^6\mathcal{P}^2}\frac{\epsilon^2}{L_{\max}} \end{aligned} \tag{B.101}$$

   where in (a) use the definition of g_{th}^2 and $\nu \geq L_{\max}/c$.

2. The gradient is small in all block directions, namely $\|\mathbf{x}^{(t+1)} - \mathbf{x}^{(t)}\|^2 \leq g_{\mathsf{th}}/\nu$: in this case, we will add the perturbation to the iterates, and implement APP for the next t_{th} steps and then check the termination condition. If the termination condition is not satisfied, we must have

$$f(\mathbf{x}^{(t_{\mathsf{th}})}) - f(\mathbf{x}^{(0)}) \leq -f_{\mathsf{th}} = -\frac{c^5\epsilon^2}{L_{\max}\chi^6\mathcal{P}^2}, \tag{B.102}$$

which implies that the objective value in each step on average is decreased by

$$\frac{f(\mathbf{x}^{(t_{\mathsf{th}})}) - f(\mathbf{x}^{(0)})}{t_{\mathsf{th}}} \leq -\frac{c^7}{\chi^7\mathcal{P}^2}\frac{\epsilon^2}{L_{\max}}\frac{(L_{\max}\rho\epsilon)^{\frac{1}{3}}}{L_{\max}}. \tag{B.103}$$

Since $\kappa = L_{\max}/(L_{\max}\rho\epsilon)^{1/3} \geq 1$ and $c \leq 1/3$, we know that RHS of (B.103) is greater than RHS of (B.101).

With the results of these two cases, we can know that if there is a large size of the gradient, we can know the decrease of the objective function value by the result of case 1, and if not, we use the result of case 2. In summary, PA-PP can have a sufficient decrease of the objective function value by $\frac{c^7}{\chi^7\mathcal{P}^2}\frac{\epsilon^2}{L_{\max}}\frac{(L_{\max}\rho\epsilon)^{1/3}}{L_{\max}}$ per iteration on average. This means that Algorithm 2 must stop within a finite number of iterations, which is

$$\frac{f(\mathbf{h}_{-1}^{(0)}, \mathbf{x}_1^{(0)}) - f^*}{\frac{c^7}{\chi^7\mathcal{P}^2}\frac{\epsilon^2}{L_{\max}}\frac{(L_{\max}\rho\epsilon)^{1/3}}{L_{\max}}} = \frac{\chi^7\mathcal{P}^2}{c^7}\frac{L_{\max}^2\Delta f}{\epsilon^2(L_{\max}\rho\epsilon)^{1/3}} = \mathcal{O}\left(\frac{\Delta f\chi^7\mathcal{P}^2L_{\max}^{5/3}}{\rho^{1/3}\epsilon^{7/3}}\right) \tag{B.104}$$

where $\Delta f := f(\mathbf{h}_{-1}^{(0)}, \mathbf{x}_1^{(0)}) - f^*$.

According to Lemma 6, we know that with probability $1 - \frac{dL_{\max}}{(L_{\max}\rho\epsilon)^{1/3}}e^{-\chi}$ the algorithm can give a sufficient descent with the perturbation when $\|\mathbf{x}^{(t+1)} - \mathbf{x}^{(t)}\|^2 \leq g_{\mathsf{th}}/\nu$. Since the total number of perturbation we can add is at most

$$n' = \frac{1}{t_{\mathsf{th}}}\frac{\chi^7\mathcal{P}^2}{c^7}\frac{L_{\max}^2\Delta f}{\epsilon^2(L_{\max}\rho\epsilon)^{1/3}} = \frac{\chi^6\mathcal{P}^2}{c^5}\frac{L_{\max}\Delta_f}{\epsilon^2}. \tag{B.105}$$

Using the union bound, the probability of Lemma 6 being satisfied for all perturbations is

$$1 - n'\frac{dL_{\max}}{(L_{\max}\rho\epsilon)^{\frac{1}{3}}}e^{-\chi} = 1 - \frac{dL_{\max}}{(L_{\max}\rho\epsilon)^{\frac{1}{3}}}e^{-\chi}\frac{\chi^6\mathcal{P}^2}{c^5}\frac{L_{\max}\Delta_f}{\epsilon^2} = 1 - \underbrace{\frac{dL_{\max}}{(L_{\max}\rho\epsilon)^{\frac{1}{3}}}\frac{\mathcal{P}^2}{c^5}\frac{\Delta_f}{\epsilon^2}}_{:=\mathcal{C}'}\chi^6e^{-\chi}. \tag{B.106}$$

With chosen $\chi = 6\max\{\ln(\mathcal{C}'/\delta), 4\}$, we have $\chi^6e^{-\chi} \leq e^{-\chi/6}$, which implies $\chi^6e^{-\chi}\mathcal{C}' \leq e^{-\chi/6}\mathcal{C}' \leq \delta$.

The proof is complete. □

B.3.2 Proof of Corollary 4

Proof. Recall the definitions:

$$\mathbf{H}'_u = \begin{bmatrix} 0 & \nabla^2_{12} f(\widetilde{\mathbf{x}}^{(t)}) \\ 0 & 0 \end{bmatrix}, \quad \mathbf{H}'_l = \begin{bmatrix} \nabla^2_{11} f(\widetilde{\mathbf{x}}^{(t)}) & 0 \\ \nabla^2_{21} f(\widetilde{\mathbf{x}}^{(t)}) & \nabla^2_{22} f(\widetilde{\mathbf{x}}^{(t)}) \end{bmatrix}, \tag{B.107}$$

where $\widetilde{\mathbf{x}}^{(t)}$ is an ϵ-second order stationary point, and

$$\mathbf{M}' := \mathbf{I} + \eta \mathbf{H}'_l \quad \mathbf{T}' := \mathbf{I} - \eta \mathbf{H}'_u. \tag{B.108}$$

Obviously, we also have $\mathbf{H} = \mathbf{H}'_l + \mathbf{H}'_u$.

Note that $\det(\mathbf{T}') = 1$, which implies that $\det(\mathbf{T}'^{-1}\mathbf{M}' - \lambda \mathbf{I}) = \det(\mathbf{M}' - \lambda \mathbf{T}')$, where λ denotes the eigenvalue. We can analyze the determinant of $\mathbf{M}' - \lambda \mathbf{T}'$. We have

$$\det[\mathbf{M}' - \lambda \mathbf{T}'] = \left[\underbrace{\begin{matrix} (1-\lambda)\mathbf{I} + \eta \nabla^2_{11} f(\widetilde{\mathbf{x}}^{(t)}) & \lambda \eta \nabla^2_{12} f(\widetilde{\mathbf{x}}^{(t)}) \\ \eta \nabla^2_{21} f(\widetilde{\mathbf{x}}^{(t)}) & (1-\lambda)\mathbf{I} + \eta \nabla^2_{22} f(\widetilde{\mathbf{x}}^{(t)}) \end{matrix}}_{:=\mathbf{Q}'(\lambda)} \right].$$

It can be observed that

$$\mathbf{Q}'(\lambda) = \begin{bmatrix} \mathbf{I} & \\ & \frac{1}{\sqrt{\lambda}} \end{bmatrix} \underbrace{\begin{bmatrix} (1-\lambda)\mathbf{I} + \eta \nabla^2_{11} f(\widetilde{\mathbf{x}}^{(t)}) & \eta\sqrt{\lambda} \nabla^2_{12} f(\widetilde{\mathbf{x}}^{(t)}) \\ \eta\sqrt{\lambda} \nabla^2_{21} f(\widetilde{\mathbf{x}}^{(t)}) & (1-\lambda)\mathbf{I} + \eta \nabla^2_{22} f(\widetilde{\mathbf{x}}^{(t)}) \end{bmatrix}}_{\mathbf{G}'(\lambda)} \begin{bmatrix} \mathbf{I} & \\ & \sqrt{\lambda} \end{bmatrix},$$

meaning that $\mathbf{Q}'(\lambda)$ is similar to $\mathbf{G}'(\lambda)$. Consequently, we can conclude that $\mathbf{Q}'(\delta)$ has the same eigenvalues of $\mathbf{G}'(\delta)$. Furthermore, since matrix $\mathbf{G}'(\lambda)$ is symmetric, we know that all eigenvalues of $\mathbf{Q}'(\lambda)$ and $\mathbf{G}'(\lambda)$ are real. Then, we can need to show there exists λ such that $\det(\mathbf{Q}'(\lambda)) = 0$.

Consider $0 \leq \delta \leq 1$. We have

$$\mathbf{G}'(1-\delta) = \begin{bmatrix} \delta \mathbf{I} + \eta \nabla^2_{11} f(\widetilde{\mathbf{x}}^{(t)}) & \eta\sqrt{1-\delta} \nabla^2_{12} f(\widetilde{\mathbf{x}}^{(t)}) \\ \eta\sqrt{1-\delta} \nabla^2_{21} f(\widetilde{\mathbf{x}}^{(t)}) & \delta \mathbf{I} + \eta \nabla^2_{22} f(\widetilde{\mathbf{x}}^{(t)}) \end{bmatrix}. \tag{B.109}$$

Since we know that $\mathbf{H}$ and $\mathbf{G}(1-\delta)$ are diagonalizable (normal matrices), then we have the following result (107) (or (108)) of quantifying the difference of the eigenvalues of the two matrices

$$\max_{1\leq i\leq d} |\lambda_i(\eta\mathbf{H}) - \lambda_i(\mathbf{G}'(1-\delta))| \leq \|\eta\mathbf{H} - \mathbf{G}'(1-\delta)\| \tag{B.110}$$

where $\lambda_i(\mathbf{H})$ and $\lambda_i(\mathbf{G}'(1-\delta))$ denote the ith eigenvalue of $\mathbf{H}$ and $\mathbf{G}'(1-\delta)$, which are listed in a decreasing order.

With the help of (B.110), we can check

$$\begin{aligned}
&\|\mathbf{G}'(1-\delta) - \eta\mathbf{H}\| \\
=&\left\| \delta\mathbf{I} + \begin{bmatrix} 0 & (\sqrt{1-\delta}-1)\eta\nabla^2_{12} f(\widetilde{\mathbf{x}}^{(t)}) \\ (\sqrt{1-\delta}-1)\eta\nabla^2_{21} f(\widetilde{\mathbf{x}}^{(t)}) & 0 \end{bmatrix} \right\| \\
\leq&\delta + (\sqrt{1-\delta}-1)\eta\|\mathbf{H}\| + (\sqrt{1-\delta}-1)\eta \left\| \begin{matrix} \nabla^2_{11} f(\widetilde{\mathbf{x}}^{(t)}) & 0 \\ 0 & \nabla^2_{22} f(\widetilde{\mathbf{x}}^{(t)}) \end{matrix} \right\| \\
\overset{(a)}{\leq}&\delta + (\sqrt{1-\delta}-1)(\frac{L}{L_{\max}}+1)
\end{aligned} \tag{B.111}$$

where (a) is true since we used $\eta \leq c_{\max}/L_{\max}$. Also, it can be observed that when $\delta = 0$, matrix $\mathbf{G}'(\delta)$ is reduced to $\eta\mathbf{H}$.

We know that the minimum eigenvalue of $\eta\mathbf{H}$ which is $-\eta\gamma$ and the maximum difference of the eigenvalues between $\eta\mathbf{H}$ and $\mathbf{G}'(\delta)$ is upper bounded by (B.111). Then, we can choose a sufficient small δ such that $\mathbf{G}'(\delta)$ also has a negative eigenvalue, meaning that we need to find a $\delta \in [0,1]$ such that

$$\delta + (\sqrt{1-\delta}-1)(\frac{L}{L_{\max}}+1) < \eta\gamma. \tag{B.112}$$

In other words, if we choose

$$\delta^* = \frac{\eta\gamma}{2}$$

then we can conclude that $\mathbf{G}'(\delta^*)$ has a negative eigenvalue which is less than $-\eta\gamma + \delta^* = -\frac{\eta\gamma}{2}$. In the following, we will check that δ^* is a valid choice, meaning that equation (B.112) holds when $\delta^* = \frac{\eta\gamma}{2}$.

Actually, equation (B.112) can be rewritten as

$$\delta + \sqrt{1-\delta}(1 + \frac{L}{L_{\max}}) < \eta\gamma + (1 + \frac{L}{L_{\max}}), \tag{B.113}$$

Since $\kappa = L_{\max}/\gamma \geq 1$ and $\eta \leq c_{\max}/L_{\max}$ where $c_{\max} \leq 1/2$, we have

$$\sqrt{1-\delta^*} = \sqrt{1-\eta\gamma/2} < 1, \tag{B.114}$$

which implies that equation (B.112) is true with chosen δ^* Therefore, we can conclude that $\mathbf{Q}'(1+\delta^*)$ has a negative eigenvalue.

When δ is large, i.e., $\delta > 1$, we have

$$\mathbf{Q}'(1-\delta) = \begin{bmatrix} \mathbf{I} & \\ & \frac{j}{\sqrt{1-\delta}} \end{bmatrix} \underbrace{\begin{bmatrix} \delta\mathbf{I} + \eta\nabla^2_{11} f(\widetilde{\mathbf{x}}^{(t)}) & -j\eta\sqrt{1-\delta}\nabla^2_{12} f(\widetilde{\mathbf{x}}^{(t)}) \\ \eta\sqrt{1-\delta}\nabla^2_{21} f(\widetilde{\mathbf{x}}^{(t)}) & \delta\mathbf{I} + \eta\nabla^2_{22} f(\widetilde{\mathbf{x}}^{(t)}) \end{bmatrix}}_{\mathbf{G}'(1-\delta)} \begin{bmatrix} \mathbf{I} & \\ & j\sqrt{1-\delta} \end{bmatrix}, \tag{B.115}$$

where j denotes the imaginary number, so $\mathbf{Q}'(1-\delta)$ is similar to $\mathbf{G}'(1-\delta)$ when $\delta > 1$. Also, we know that $\mathbf{G}'(1-\delta)$ is a Hermitian matrix. It is easy to check $\mathbf{Q}'(1-\delta)$ has a positive eigenvalue, since term $\delta\mathbf{I}$ dominates the spectrum of matrix $\mathbf{Q}'(1-\delta)$ in (B.115). Considering the eigenvalue is continuous with respect to δ, we can conclude there exists a δ, i.e., $\widehat{\delta}'$, such that $\mathbf{Q}'(1-\widehat{\delta}')$ has a zero eigenvalue, i.e., $\det(\mathbf{Q}'(1-\widehat{\delta}')) = 0$ where $1-\widehat{\delta}'$ is at least as small as

$$1 - \delta^* = 1 - \frac{\eta\gamma}{2}, \tag{B.116}$$

meaning that $1-\widehat{\delta}' \leq 1 - \frac{\eta\gamma}{2}$. □

In the following, we will give the proofs of Lemma 18–Lemma 22 in details.

Proofs of Lemma 17–Lemma 22

B.3.3 Proof of Lemma 17

Proof. First, we have

$$\begin{aligned}\|\nabla_1 f(\mathbf{x}_1^{(t)}, \mathbf{x}_2^{(t)})\|^2 \leq & 2\|\nabla_1 f(\mathbf{x}_1^{(t+1)}, \mathbf{x}_2^{(t)}) - \nabla_1 f(\mathbf{x}_1^{(t)}, \mathbf{x}_2^{(t)})\|^2 + 2\|\nabla_1 f(\mathbf{x}_1^{(t+1)}, \mathbf{x}_2^{(t)})\|^2 \\ \overset{(a)}{\leq} & 2L_{\max}^2 \|\mathbf{x}_1^{(t+1)} - \mathbf{x}_1^{(t)}\|^2 + 2\|\nabla_1 f(\mathbf{x}_1^{(t+1)}, \mathbf{x}_2^{(t)})\|^2 \\ \overset{(4.7)}{\leq} & 2L_{\max}^2 \eta^2 \|\nabla_1 f(\mathbf{x}_1^{(t+1)}, \mathbf{x}_2^{(t)})\|^2 + 2\|\nabla_1 f(\mathbf{x}_1^{(t+1)}, \mathbf{x}_2^{(t)})\|^2 \\ \overset{(b)}{\leq} & 3\|\nabla_1 f(\mathbf{x}_1^{(t+1)}, \mathbf{x}_2^{(t)})\|^2 \end{aligned} \tag{B.117}$$

where in (a) we used block-wise Lipschitz continuity, in (b) we choose $\eta \leq 1/(2L_{\max})$.

$$\begin{aligned}\|\nabla_2 f(\mathbf{x}_1^{(t)}, \mathbf{x}_2^{(t)})\|^2 \leq & 2\|\nabla_2 f(\mathbf{x}_1^{(t+1)}, \mathbf{x}_2^{(t+1)}) - \nabla_2 f(\mathbf{x}_1^{(t)}, \mathbf{x}_2^{(t)})\|^2 + 2\|\nabla_2 f(\mathbf{x}_1^{(t+1)}, \mathbf{x}_2^{(t+1)})\|^2 \\ \leq & 4(\|\nabla_2 f(\mathbf{x}_1^{(t+1)}, \mathbf{x}_2^{(t+1)}) - \nabla_2 f(\mathbf{x}_1^{(t+1)}, \mathbf{x}_2^{(t)})\|^2 \\ & + \|\nabla_2 f(\mathbf{x}_1^{(t+1)}, \mathbf{x}_2^{(t)}) - \nabla_2 f(\mathbf{x}_1^{(t)}, \mathbf{x}_2^{(t)})\|^2) + 2\|\nabla_2 f(\mathbf{x}_1^{(t+1)}, \mathbf{x}_2^{(t+1)})\|^2 \\ \overset{(4.7)}{\leq} & 4(L_{\max}^2 \|\mathbf{x}_2^{(t+1)} - \mathbf{x}_2^{(t)}\|^2 + \|\mathbf{x}_1^{(t+1)} - \mathbf{x}_1^{(t)}\|^2) + 2\|\nabla_2 f(\mathbf{x}_1^{(t+1)}, \mathbf{x}_2^{(t+1)})\|^2 \\ \overset{(a)}{\leq} & \|\nabla_1 f(\mathbf{x}_1^{(t+1)}, \mathbf{x}_2^{(t)})\|^2 + 3\|\nabla_2 f(\mathbf{x}_1^{(t+1)}, \mathbf{x}_2^{(t+1)})\|^2 \end{aligned} \tag{B.118}$$

where (a) we also choose $\eta \leq 1/(2L_{\max})$.

Summing (B.117) and (B.118), we have

$$\|\nabla f(\mathbf{x}^{(t)})\|^2 \leq \sum_{k=1}^{2} \|\nabla_k f(\mathbf{x}_k^{(t)})\|^2 \leq 4 \sum_{k=1}^{2} \|\nabla_k f(\mathbf{h}_{-k}^{(t)}, \mathbf{x}_k^{(t+1)})\|^2 \overset{(4.7)}{=} 4\nu \|\mathbf{x}^{(t+1)} - \mathbf{x}^{(t)}\|^2 \tag{B.119}$$

where $\mathbf{h}_{-1}^{(t)} = \mathbf{x}_2^{(t)}$ and $\mathbf{h}_{-2}^{(t)} = \mathbf{x}_1^{(t+1)}$. □

B.3.4 Proof of Lemma 18

There proof involves two parts:

Upper Triangular Matrix: Consider three different vectors $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$. We can have

$$\begin{aligned}
&\left\| \begin{bmatrix} \nabla^2_{11} f(\mathbf{x}) & 0 \\ \nabla^2_{21} f(\mathbf{y}) & \nabla^2_{22} f(\mathbf{y}) \end{bmatrix} - \begin{bmatrix} \nabla^2_{11} f(\mathbf{z}) & 0 \\ \nabla^2_{21} f(\mathbf{z}) & \nabla^2_{22} f(\mathbf{z}) \end{bmatrix} \right\| \\
\leq &\left\| \mathbf{I}_1 \left(\begin{bmatrix} \nabla^2_{11} f(\mathbf{x}) & \nabla^2_{12} f(\mathbf{x}) \\ \nabla^2_{21} f(\mathbf{x}) & \nabla^2_{22} f(\mathbf{x}) \end{bmatrix} - \begin{bmatrix} \nabla^2_{11} f(\mathbf{z}) & \nabla^2_{12} f(\mathbf{z}) \\ \nabla^2_{21} f(\mathbf{z}) & \nabla^2_{22} f(\mathbf{z}) \end{bmatrix} \right) \mathbf{I}_1 \right\| \\
&+ \left\| \mathbf{I}_2 \left(\begin{bmatrix} \nabla^2_{11} f(\mathbf{y}) & \nabla^2_{12} f(\mathbf{y}) \\ \nabla^2_{21} f(\mathbf{y}) & \nabla^2_{22} f(\mathbf{y}) \end{bmatrix} - \begin{bmatrix} \nabla^2_{11} f(\mathbf{z}) & \nabla^2_{12} f(\mathbf{z}) \\ \nabla^2_{21} f(\mathbf{z}) & \nabla^2_{22} f(\mathbf{z}) \end{bmatrix} \right) \right\| \\
\overset{(a)}{\leq} &\left\| \begin{bmatrix} \nabla^2_{11} f(\mathbf{x}) & \nabla^2_{12} f(\mathbf{x}) \\ \nabla^2_{21} f(\mathbf{x}) & \nabla^2_{22} f(\mathbf{x}) \end{bmatrix} - \begin{bmatrix} \nabla^2_{11} f(\mathbf{z}) & \nabla^2_{12} f(\mathbf{z}) \\ \nabla^2_{21} f(\mathbf{z}) & \nabla^2_{22} f(\mathbf{z}) \end{bmatrix} \right\| \\
&+ \left\| \begin{bmatrix} \nabla^2_{11} f(\mathbf{y}) & \nabla^2_{12} f(\mathbf{y}) \\ \nabla^2_{21} f(\mathbf{y}) & \nabla^2_{22} f(\mathbf{y}) \end{bmatrix} - \begin{bmatrix} \nabla^2_{11} f(\mathbf{z}) & \nabla^2_{12} f(\mathbf{z}) \\ \nabla^2_{21} f(\mathbf{z}) & \nabla^2_{22} f(\mathbf{z}) \end{bmatrix} \right\| \\
\leq &\rho \left(\|\mathbf{x} - \mathbf{z}\| + \|\mathbf{y} - \mathbf{z}\| \right)
\end{aligned}$$

where in (a) we use

$$\mathbf{I}_1 = \begin{bmatrix} \mathbf{I} & 0 \\ 0 & 0 \end{bmatrix} \qquad \mathbf{I}_2 = \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{I} \end{bmatrix} \tag{B.120}$$

and $\|\mathbf{I}_1\| = \|\mathbf{I}_2\| = 1$.

Lower Triangular Matrix:

$$\begin{aligned}
&\left\| \begin{bmatrix} 0 & \nabla^2_{21} f(\mathbf{x}) \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} 0 & \nabla^2_{21} f(\mathbf{y}) \\ 0 & 0 \end{bmatrix} \right\| \\
= &\left\| \mathbf{I}_1 \left(\begin{bmatrix} \nabla^2_{11} f(\mathbf{x}) & \nabla^2_{12} f(\mathbf{x}) \\ \nabla^2_{21} f(\mathbf{x}) & \nabla^2_{22} f(\mathbf{x}) \end{bmatrix} - \begin{bmatrix} \nabla^2_{11} f(\mathbf{y}) & \nabla^2_{12} f(\mathbf{y}) \\ \nabla^2_{21} f(\mathbf{y}) & \nabla^2_{22} f(\mathbf{y}) \end{bmatrix} \right) \mathbf{I}_2 \right\| \\
\overset{(a)}{\leq} &\rho \|\mathbf{x} - \mathbf{y}\|
\end{aligned}$$

where (a) is true because we know $\|\mathbf{I}_1\| = \|\mathbf{I}_2\| = 1$.

B.3.5 Proof of Lemma 19

Proof. Under Assumption 1, we have (descent lemma)

$$\begin{aligned}
f(\mathbf{x}^{(t+1)}) \leq & f(\mathbf{x}^{(t)}) + \sum_{k=1}^{2} \nabla_k f(\mathbf{h}_{-k}^{(t)}, \mathbf{x}_k^{(t)})^T (\mathbf{x}_k^{(t+1)} - \mathbf{x}_k^{(t)}) + \sum_{k=1}^{2} \frac{L_k}{2} \|\mathbf{x}_k^{(t+1)} - \mathbf{x}_k^{(t)}\|^2 \\
\leq & f(\mathbf{x}^{(t)}) + \sum_{k=1}^{2} \nabla_k f(\mathbf{h}_{-k}^{(t)}, \mathbf{x}_k^{(t+1)})^T (\mathbf{x}_k^{(t+1)} - \mathbf{x}_k^{(t)}) \\
& + \sum_{k=1}^{2} (\nabla_k f(\mathbf{h}_{-k}^{(t)}, \mathbf{x}_k^{(t)}) - \nabla_k f(\mathbf{h}_{-k}^{(t)}, \mathbf{x}_k^{(t+1)}))^T (\mathbf{x}_k^{(t+1)} - \mathbf{x}_k^{(t)}) + \sum_{k=1}^{2} \frac{L_k}{2} \|\mathbf{x}_k^{(t+1)} - \mathbf{x}_k^{(t)}\|^2 \\
\overset{(a)}{\leq} & f(\mathbf{x}^{(t)}) - \sum_{k=1}^{2} \eta \|\nabla_k f(\mathbf{h}_{-k}^{(t)}, \mathbf{x}_k^{(t+1)})\|^2 + \sum_{k=1}^{2} \frac{3\eta^2 L_k}{2} \|\nabla_k f(\mathbf{h}_{-k}^{(t)}, \mathbf{x}_k^{(t+1)})\|^2 \\
\overset{(b)}{\leq} & f(\mathbf{x}^{(t+1)}) - \sum_{k=1}^{2} \frac{\eta}{2} \|\nabla_k f(\mathbf{h}_{-k}^{(t)}, \mathbf{x}_k^{(t+1)})\|^2 \\
= & f(\mathbf{x}^{(t+1)}) - \frac{\nu}{2} \|\mathbf{x}^{(t+1)} - \mathbf{x}^{(t)}\|^2
\end{aligned} \tag{B.121}$$

where (a) is true because of the update rule of APP in each block and Assumption 1 and block-wise Lipschitz continuity, in (b) we choose $\eta \leq 1/(3L_{\max})$ and $\nu = 1/\eta$. □

B.3.6 Proof of Lemma 20

Proof. Without loss of generality, let $\mathbf{u}^{(0)}$ be the origin, i.e., $\mathbf{u}^{(0)} = 0$. According to the APP update rule of variables, we have

$$\mathbf{u}^{(t+1)} = \mathbf{u}^{(t)} - \eta \begin{bmatrix} \nabla_1 f(\mathbf{u}_1^{(t+1)}, \mathbf{u}_2^{(t)}) \\ \nabla_2 f(\mathbf{u}_1^{(t+1)}, \mathbf{u}_2^{(t+1)}) \end{bmatrix}. \tag{B.122}$$

It can be observed that the update rule of PA-PP is very similar as the one of PA-GD. The proof of this lemma is also similar as Lemma 14. We only need to replace $\nabla_1 f(\mathbf{u}_1^{(t)}, \mathbf{u}_2^{(t)})$ as $\nabla_1 f(\mathbf{u}_1^{(t+1)}, \mathbf{u}_2^{(t)})$ and $\nabla_2 f(\mathbf{u}_1^{(t+1)}, \mathbf{u}_2^{(t)})$ as $\nabla_2 f(\mathbf{u}_1^{(t+1)}, \mathbf{u}_2^{(t+1)})$, which can give us the claimed result after following the proof of Lemma 14. Hence, we ignore the repeated part with the proof of Lemma 14 for simplicity of expressions. □

B.3.7 Proof of Lemma 21

Proof. Let $\mathbf{u}^{(0)} = 0$ and define $\mathbf{v}^{(t)} := \mathbf{w}^{(t)} - \mathbf{u}^{(t)}$. According to the assumption of Lemma 15, we know that $\mathbf{v}^{(0)} = \upsilon[\eta L_{\max}\mathcal{S}/(\kappa \log(\frac{d\kappa}{\delta})\mathcal{P}_1)]\vec{\mathbf{e}}'$ when $\upsilon \in [\delta/(2\sqrt{d}), 1]$. First, we define the following auxiliary function

$$h(\theta) := \begin{bmatrix} \nabla_1 f(\mathbf{u}_1^{(t+1)} + \theta\mathbf{v}_1^{(t+1)}, \mathbf{u}_2^{(t)} + \theta\mathbf{v}_2^{(t)}) \\ \nabla_2 f(\mathbf{u}_1^{(t+1)} + \theta\mathbf{v}_1^{(t+1)}, \mathbf{u}_2^{(t+1)} + \theta\mathbf{v}_2^{(t+1)}) \end{bmatrix},$$

then have

$$h(0) = \begin{bmatrix} \nabla_1 f(\mathbf{u}_1^{(t+1)}, \mathbf{u}_2^{(t)}) \\ \nabla_2 f(\mathbf{u}_1^{(t+1)}, \mathbf{u}_2^{(t+1)}) \end{bmatrix}, \quad h(1) = \begin{bmatrix} \nabla_1 f(\mathbf{u}_1^{(t+1)} + \mathbf{v}_1^{(t+1)}, \mathbf{u}_2^{(t)} + \mathbf{v}_2^{(t)}) \\ \nabla_2 f(\mathbf{u}_1^{(t+1)} + \mathbf{v}_1^{(t+1)}, \mathbf{u}_2^{(t+1)} + \mathbf{v}_2^{(t+1)}) \end{bmatrix},$$

$$g(\theta) = \frac{dh(\theta)}{d\theta}$$

$$= \underbrace{\begin{bmatrix} \nabla_{11}^2 f(\mathbf{u}_1^{(t+1)} + \theta\mathbf{v}_1^{(t+1)}, \mathbf{u}_2^{(t)} + \theta\mathbf{v}_2^{(t)}) & 0 \\ \nabla_{21}^2 f(\mathbf{u}_1^{(t+1)} + \theta\mathbf{v}_1^{(t+1)}, \mathbf{u}_2^{(t+1)} + \theta\mathbf{v}_2^{(t+1)}) & \nabla_{22}^2 f(\mathbf{u}_1^{(t+1)} + \theta\mathbf{v}_1^{(t+1)}, \mathbf{u}_2^{(t+1)} + \theta\mathbf{v}_2^{(t+1)}) \end{bmatrix}}_{\widetilde{\boldsymbol{\mathcal{H}}}_l'^{(t)}(\theta)} \mathbf{v}^{(t+1)}$$

$$+ \underbrace{\begin{bmatrix} 0 & \nabla_{12}^2 f(\mathbf{u}_1^{(t+1)} + \theta\mathbf{v}_1^{(t+1)}, \mathbf{u}_2^{(t)} + \theta\mathbf{v}_2^{(t)}) \\ 0 & 0 \end{bmatrix}}_{\widetilde{\boldsymbol{\mathcal{H}}}_u'^{(t)}(\theta)} \mathbf{v}^{(t)},$$

$$\begin{bmatrix} \nabla_1 f(\mathbf{w}_1^{(t+1)}, \mathbf{w}_2^{(t)}) \\ \nabla_2 f(\mathbf{w}_1^{(t+1)}, \mathbf{w}_2^{(t+1)}) \end{bmatrix} = \int_0^1 g(\theta)d\theta + \begin{bmatrix} \nabla_1 f(\mathbf{u}_1^{(t+1)}, \mathbf{u}_2^{(t)}) \\ \nabla_2 f(\mathbf{u}_1^{(t+1)}, \mathbf{u}_2^{(t+1)}) \end{bmatrix}.$$

Then, we consider sequence $\mathbf{w}^{(t)}$, i.e.,

$$\mathbf{u}^{(t+1)}+\mathbf{v}^{(t+1)}=\mathbf{w}^{(t+1)}=\mathbf{w}^{(t)}-\eta\begin{bmatrix}\nabla_1 f(\mathbf{w}_1^{(t+1)},\mathbf{w}_2^{(t)})\\ \nabla_2 f(\mathbf{w}_1^{(t+1)},\mathbf{w}_2^{(t+1)})\end{bmatrix} \tag{B.123}$$

$$\begin{aligned}=&\mathbf{u}^{(t)}+\mathbf{v}^{(t)}-\eta\begin{bmatrix}\nabla_1 f(\mathbf{u}_1^{(t+1)}+\mathbf{v}_1^{(t+1)},\mathbf{u}_1^{(t)}+\mathbf{v}_1^{(t)})\\ \nabla_2 f(\mathbf{u}_1^{(t+1)}+\mathbf{v}_1^{(t+1)},\mathbf{u}_2^{(t+1)}+\mathbf{v}_2^{(t+1)})\end{bmatrix}\\ =&\mathbf{u}^{(t)}+\mathbf{v}^{(t)}-\eta\begin{bmatrix}\nabla_1 f(\mathbf{u}_1^{(t+1)},\mathbf{u}_2^{(t)})\\ \nabla_2 f(\mathbf{u}_1^{(t+1)},\mathbf{u}_2^{(t+1)})\end{bmatrix}-\int_0^1 g(\theta)d\theta \qquad \text{(B.124)}\\ \overset{(a)}{=}&\mathbf{u}^{(t)}+\mathbf{v}^{(t)}-\eta\begin{bmatrix}\nabla_1 f(\mathbf{u}_1^{(t+1)},\mathbf{u}_2^{(t)})\\ \nabla_2 f(\mathbf{u}_1^{(t+1)},\mathbf{u}_2^{(t+1)})\end{bmatrix}-\eta\widetilde{\Delta}_u'^{(t)}\mathbf{v}^{(t)}-\mathbf{H}_u'\mathbf{v}^{(t)}-\eta\widetilde{\Delta}_l'^{(t)}\mathbf{v}^{(t+1)}-\eta\mathbf{H}_l'\mathbf{v}^{(t+1)}\end{aligned} \tag{B.125}$$

where in (a) we used the following definitions

$$\widetilde{\Delta}_u'^{(t)}:=\int_0^1\widetilde{\boldsymbol{\mathcal{H}}}_u'^{(t)}(\theta)d\theta-\mathbf{H}_u',$$
$$\widetilde{\Delta}_l'^{(t)}:=\int_0^1\widetilde{\boldsymbol{\mathcal{H}}}_l'^{(t)}(\theta)d\theta-\mathbf{H}_l',$$

and

$$\mathbf{H}_u'=\begin{bmatrix}0&\nabla_{12}^2 f(\widetilde{\mathbf{x}}^{(t)})\\ 0&0\end{bmatrix}\quad\mathbf{H}_l'=\begin{bmatrix}\nabla_{11}^2 f(\widetilde{\mathbf{x}}^{(t)})&0\\ \nabla_{21}^2 f(\widetilde{\mathbf{x}}^{(t)})&\nabla_{22}^2 f(\widetilde{\mathbf{x}}^{(t)})\end{bmatrix}. \tag{B.126}$$

Obviously, $\mathbf{H}=\mathbf{H}_l'+\mathbf{H}_u'$.

Dynamics of $\mathbf{v}^{(t)}$: Since the first two terms at RHS of (B.125) combined with $\mathbf{u}^{(t)}$ at LHS of (B.125) are exactly the same as (B.122). It can be observed that equation (B.125) gives the dynamic of $\mathbf{v}^{(t)}$, i.e.,

$$\mathbf{v}^{(t+1)}=\mathbf{v}^{(t)}-\eta\widetilde{\Delta}_u'^{(t)}\mathbf{v}^{(t)}-\eta\mathbf{H}_u'\mathbf{v}^{(t)}-\eta\widetilde{\Delta}_l'^{(t)}\mathbf{v}^{(t+1)}-\eta\mathbf{H}_l'\mathbf{v}^{(t+1)}, \tag{B.127}$$

which can be equivalently expressed by

$$\underbrace{(\mathbf{I}+\eta\mathbf{H}_l')}_{:=\mathbf{M}'}\mathbf{v}^{(t+1)}=\underbrace{(\mathbf{I}-\eta\mathbf{H}_u')}_{:=\mathbf{T}'}\mathbf{v}^{(t)}-\eta\widetilde{\Delta}_l'^{(t)}\mathbf{v}^{(t+1)}-\eta\widetilde{\Delta}_u'^{(t)}\mathbf{v}^{(t)}. \tag{B.128}$$

It is worth noting that matrix $\mathbf{T}'$ is an upper triangular matrix where the diagonal entries are all 1s, so it is invertible. Taking the inverse of $\mathbf{T}'$ on both sides of (B.128), we can obtain

$$\mathbf{T}'^{-1}\mathbf{M}'\mathbf{v}^{(t+1)} \overset{(B.125)}{=} \mathbf{v}^{(t)} - \mathbf{T}'^{-1}\eta\widetilde{\Delta}_l'^{(t)}\mathbf{v}^{(t+1)} - \mathbf{T}'^{-1}\eta\widetilde{\Delta}_u'^{(t)}\mathbf{v}^{(t)}. \tag{B.129}$$

Let $\mathbb{P}'_{\texttt{left}}$ denote the projection operator that projects the vector onto the space spanned by the eigenvector of $\mathbf{T}'^{-1}\mathbf{M}$ whose corresponding positive eigenvalue is minimum. Taking the projection on both sides of (B.129), we have

$$\mathbb{P}'_{\texttt{left}}(\mathbf{T}'^{-1}\mathbf{M}')\mathbf{v}^{(t+1)} + \mathbb{P}'_{\texttt{left}}\mathbf{T}'^{-1}\eta\widetilde{\Delta}_l'^{(t)}\mathbf{v}^{(t+1)} = \mathbb{P}'_{\texttt{left}}\mathbf{v}^{(t)} - \mathbb{P}'_{\texttt{left}}\mathbf{T}'^{-1}\eta\widetilde{\Delta}_u'^{(t)}\mathbf{v}^{(t)}. \tag{B.130}$$

Relationship of the Norm of $\mathbf{v}^{(t)}$ Projected onto the Two Subspaces: Let $\phi^{(t)}$ denote the norm of $\mathbf{v}^{(t)}$ projected onto the space spanned by the eigenvector of $\mathbf{T}'^{-1}\mathbf{M}'$ whose positive minimum eigenvalue of $\mathbf{M}'^{-1}\mathbf{T}'$ is $1-\widehat{\delta}' > 0$ and $\theta^{(t)}$ denote the norm of $\mathbf{v}^{(t)}$ projected onto the remaining space. From (B.130), we can have

$$(1-\widehat{\delta}')\phi^{(t+1)} \overset{(a)}{\geq} \phi^{(t)} - \eta\|\mathbf{T}'^{-1}\|\|\widetilde{\Delta}_l'^{(t)}\|\|\mathbf{v}^{(t+1)}\| - \eta\|\mathbf{T}'^{-1}\|\|\widetilde{\Delta}_u'^{(t)}\|\|\mathbf{v}^{(t)}\|, \tag{B.131}$$

$$(1-\widehat{\delta}')\theta^{(t+1)} \leq \theta^{(t)} + \eta\|\mathbf{T}'^{-1}\|\|\widetilde{\Delta}_l'^{(t)}\|\|\mathbf{v}^{(t+1)}\| + \eta\|\mathbf{T}'^{-1}\|\|\widetilde{\Delta}_u'^{(t)}\|\|\mathbf{v}^{(t)}\|. \tag{B.132}$$

where (a) is true because we applied the triangle inequality since η is sufficiently small.

Since $\|\mathbf{w}^{(0)} - \widetilde{\mathbf{x}}^{(t)}\| \leq \|\mathbf{u}^{(0)} - \widetilde{\mathbf{x}}^{(t)}\| + \|\mathbf{v}^{(0)}\| \leq 2r$, we can apply Lemma 20. Then, we know $\|\mathbf{w}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\| \leq 5\widehat{c}\mathcal{S}, \forall t < T$. According to the assumptions of Lemma 21, we have $\|\mathbf{u}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\| \leq 5\widehat{c}\mathcal{S}$, and

$$\|\mathbf{v}^{(t)}\| = \|\mathbf{w}^{(t)} - \mathbf{u}^{(t)}\| \leq \|\mathbf{u}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\| + \|\mathbf{w}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\| \leq 10\widehat{c}\mathcal{S}. \tag{B.133}$$

From (B.41), we know that

$$\|\mathbf{w}^{(t+1)} - \mathbf{w}^{(t)}\| \leq \frac{4.3\eta\mathcal{G}}{\kappa} = \frac{4.3\eta^3 L_{\max}^3 \frac{\gamma}{\rho}}{\kappa^2 \log^3 \frac{d\kappa}{\delta}\mathcal{P}} \leq \mathcal{S},$$

where we choose $\eta \leq c_{\max}/L_{\max}$ and $c_{\max} = 1/10$. Similarly, we also have $\|\mathbf{u}^{(t+1)} - \mathbf{u}^{(t)}\| \leq \mathcal{S}$.

Then, we need to quantify the upper bounds of $\|\mathbf{M}'^{-1}\|$, $\|\mathbf{v}^{(t+1)}\|$, $\|\widetilde{\Delta}_u'^{(t)}\|$ and $\|\widetilde{\Delta}_l'^{(t)}\|$.

1. Upper bound of $\|\mathbf{M}'^{-1}\|$: applying the steps of deriving (B.57), we can quantify the inverse of matrix $\mathbf{T}'$ as follows

$$\begin{aligned}\|\mathbf{T}'^{-1}\| \leq & 1+\eta\|\mathbf{H}'_u\| = 1+\eta\|\mathbf{H}'^T_u\| \\ = & 1+\|\eta\mathbf{H}\odot\mathbf{D}-\eta\mathbf{H}_d\| \\ < & 2(1+\frac{L\log(2d)}{L_{\max}}).\end{aligned}$$

2. Relation between $\|\mathbf{v}^{(t)}\|$ and $\|\mathbf{v}^{(t+1)}\|$: We also know that

$$\begin{aligned}\|\mathbf{v}^{(t+1)}\|^2 = & \|\mathbf{w}^{(t+1)}-\mathbf{u}^{(t+1)}\|^2 \\ = & \left\|\mathbf{w}^{(t)}-\eta\begin{bmatrix}\nabla_1 f(\mathbf{w}_1^{(t+1)},\mathbf{w}_2^{(t)})\\ \nabla_2 f(\mathbf{w}_1^{(t+1)},\mathbf{w}_2^{(t+1)})\end{bmatrix}-\left(\mathbf{u}^{(t)}-\eta\begin{bmatrix}\nabla_1 f(\mathbf{u}_1^{(t+1)},\mathbf{u}_2^{(t)})\\ \nabla_2 f(\mathbf{u}_1^{(t+1)},\mathbf{u}_2^{(t+1)})\end{bmatrix}\right)\right\|^2 \\ \leq & 2\|\mathbf{v}^{(t)}\|^2+4\eta^2\left\|\begin{bmatrix}\nabla_1 f(\mathbf{w}_1^{(t+1)},\mathbf{w}_2^{(t)})\\ \nabla_2 f(\mathbf{w}_1^{(t+1)},\mathbf{w}_2^{(t+1)})\end{bmatrix}-\begin{bmatrix}\nabla_1 f(\mathbf{u}_1^{(t+1)},\mathbf{w}_2^{(t)})\\ \nabla_2 f(\mathbf{u}_1^{(t+1)},\mathbf{w}_2^{(t+1)})\end{bmatrix}\right\|^2 \\ & +4\eta^2\left\|\begin{bmatrix}\nabla_1 f(\mathbf{u}_1^{(t+1)},\mathbf{w}_2^{(t)})\\ \nabla_2 f(\mathbf{u}_1^{(t+1)},\mathbf{w}_2^{(t+1)})\end{bmatrix}-\begin{bmatrix}\nabla_1 f(\mathbf{u}_1^{(t+1)},\mathbf{u}_2^{(t)})\\ \nabla_2 f(\mathbf{u}_1^{(t+1)},\mathbf{u}_2^{(t+1)})\end{bmatrix}\right\|^2 \\ \overset{(a)}{\leq} & 2\|\mathbf{v}^{(t)}\|^2+8\eta^2L_{\max}^2\|\mathbf{v}_1^{(t)}\|^2+4\eta^2L_{\max}^2(\|\mathbf{v}_2^{(t+1)}\|^2+\|\mathbf{v}_2^{(t)}\|^2)\end{aligned} \tag{B.134}$$

where (a) is true due to Lipschitz continuity.

We can express (B.134) as

$$(1-4\eta^2L_{\max}^2)\|\mathbf{v}^{(t+1)}\| \leq (2+8\eta^2L_{\max}^2)\|\mathbf{v}^{(t)}\|^2,$$

which implies

$$\|\mathbf{v}^{(t+1)}\| \leq \sqrt{\frac{2+\frac{8}{100}}{1-\frac{4}{100}}}\|\mathbf{v}^{(t)}\| < \sqrt{2.2}\|\mathbf{v}^{(t)}\| < 1.5\|\mathbf{v}^{(t)}\| \tag{B.135}$$

where we choose $\eta \leq c_{\max}/L_{\max}$ and $c_{\max} = 1/10$.

3. Upper bound of $\|\widetilde{\Delta}_l^{\prime(t)}\|$: applying Lemma 18, we can also get the upper bound of $\|\widetilde{\Delta}_l^{\prime(t)}\|$, i.e.,

$$\begin{aligned}
\|(\widetilde{\Delta}_l^{\prime(t)})\| \leq & \int_0^1 \|\widetilde{\boldsymbol{\mathcal{H}}}_l^{\prime(t)}(\theta) - \mathbf{H}_l'\| d\theta \\
\overset{(B.90)}{\leq} & \int_0^1 \rho \left(\|\mathbf{u}^{(t+1)} + \theta \mathbf{v}^{(t+1)} - \widetilde{\mathbf{x}}^{(t)}\| + \left\| \begin{bmatrix} \mathbf{u}_1^{(t+1)} + \theta \mathbf{v}_1^{(t+1)} \\ \mathbf{u}_2^{(t)} + \theta \mathbf{v}_2^{(t)} \end{bmatrix} - \widetilde{\mathbf{x}}^{(t)} \right\| \right) d\theta \\
\leq & \int_0^1 \rho \left(2\|\mathbf{u}^{(t+1)} + \theta \mathbf{v}^{(t+1)} - \widetilde{\mathbf{x}}^{(t)}\| + \|\mathbf{u}^{(t)} + \theta \mathbf{v}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\| \right) d\theta \\
\leq & \rho(2\|\mathbf{u}^{(t+1)} - \widetilde{\mathbf{x}}^{(t)}\| + \|\mathbf{u}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\|) + \rho \int_0^1 \theta(2\|\mathbf{v}^{(t+1)}\| + \|\mathbf{v}^{(t)}\|) d\theta \\
\leq & \rho \left(2\|\mathbf{u}^{(t+1)} - \mathbf{u}^{(t)}\| + 2\|\mathbf{u}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\| + \|\mathbf{u}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\|) + 0.5\|\mathbf{v}^{(t+1)}\| + 0.5\|\mathbf{v}^{(t)}\| \right) \\
\overset{(B.135)}{\leq} & \rho \left(2\|\mathbf{u}^{(t+1)} - \mathbf{u}^{(t)}\| + 3\|\mathbf{u}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\| + 1.25\|\mathbf{v}^{(t)}\| \right) \\
\leq & \rho(2 + 27.5\widehat{c})\mathcal{S}.
\end{aligned}$$

4. Upper bound of $\|\widetilde{\Delta}_u^{\prime(t)}\|$: according to ρ-Hessian Lipschitz continuity and Lemma 18, we have the size of $\widetilde{\Delta}_u^{\prime(t)}$ as the following.

$$\begin{aligned}
\|(\widetilde{\Delta}_u^{\prime(t)})\| \leq & \int_0^1 \|\widetilde{\boldsymbol{\mathcal{H}}}_u^{\prime(t)}(\theta) - \mathbf{H}_u'\| d\theta \\
\overset{(B.91)}{\leq} & \int_0^1 \rho \left\| \begin{bmatrix} \mathbf{u}_1^{(t+1)} + \theta \mathbf{v}_1^{(t+1)} \\ \mathbf{u}_2^{(t)} + \theta \mathbf{v}_2^{(t)} \end{bmatrix} - \widetilde{\mathbf{x}}^{(t)} \right\| d\theta \\
\leq & \int_0^1 \rho(\|\mathbf{u}^{(t)} + \theta \mathbf{v}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\| + \|\mathbf{u}^{(t+1)} + \theta \mathbf{v}^{(t+1)} - \widetilde{\mathbf{x}}^{(t)}) \| d\theta \\
\leq & \rho(\|\mathbf{u}^{(t+1)} - \widetilde{\mathbf{x}}^{(t)}\| + \|\mathbf{u}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\|) + \rho \int_0^1 \theta(\|\mathbf{v}^{(t+1)}\| + \|\mathbf{v}^{(t)}\|) d\theta \\
\overset{(B.135)}{\leq} & \rho \left(\|\mathbf{u}^{(t+1)} - \mathbf{u}^{(t)}\| + 2\|\mathbf{u}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\| + 1.25\|\mathbf{v}^{(t)}\| \right) \\
\leq & \rho(1 + 22.5\widehat{c})\mathcal{S}.
\end{aligned} \tag{B.136}$$

With the bounds of $\|\mathbf{v}^{(t+1)}\|$, $\|\widetilde{\Delta}_u^{\prime(t)}\|$, $\|\widetilde{\Delta}_l^{\prime(t)}\|$ and relation between $\|\mathbf{v}^{(t+1)}\|$ and $\|\mathbf{v}^{(t)}\|$, we can further simply (B.131) and (B.132) as follows,

$$\begin{aligned}
(1 - \widehat{\delta}')\phi^{(t+1)} & \overset{(B.131)}{\geq} \phi^{(t)} - \eta(1.5\|\widetilde{\Delta}_l^{\prime(t)}\| + \|\widetilde{\Delta}_u^{\prime(t)}\|)\|\mathbf{T}'^{-1}\| \sqrt{(\phi^{(t)})^2 + (\theta^{(t)})^2}, \\
(1 - \widehat{\delta}')\theta^{(t+1)} & \overset{(B.132)}{\leq} \theta^{(t)} + \eta(1.5\|\widetilde{\Delta}_l^{\prime(t)}\| + \|\widetilde{\Delta}_u^{\prime(t)}\|)\|\mathbf{T}'^{-1}\| \sqrt{(\phi^{(t)})^2 + (\theta^{(t)})^2},
\end{aligned}$$

since $\|\mathbf{v}^{(t)}\| = \sqrt{(\phi^{(t)})^2 + (\theta^{(t)})^2}$.

Consequently, we can arrive at

$$(1 - \widehat{\delta}')\phi^{(t+1)} \geq \phi^{(t)} - \mu\sqrt{(\phi^{(t)})^2 + (\theta^{(t)})^2}, \tag{B.137}$$

$$(1 - \widehat{\delta}')\theta^{(t+1)} \leq \theta^{(t)} + \mu\sqrt{(\phi^{(t)})^2 + (\theta^{(t)})^2}, \tag{B.138}$$

where μ is the upper bound of term $\eta(1.5\|\widetilde{\Delta}_l'^{(t)}\| + \|\widetilde{\Delta}_u'^{(t)}\|)\|\mathbf{T}'^{-1}\|$ and can be obtained by

$$\mu := \eta\rho\mathcal{S}\mathcal{P}(4 + 62\widehat{c}). \tag{B.139}$$

Quantifying the Norm of $\mathbf{v}^{(t)}$ Projected at Different Subspaces: Then, we will use mathematical induction to prove

$$\theta^{(t)} \leq 4\mu t\phi^{(t)}. \tag{B.140}$$

It is true when $t = 0$ since $\|\theta^{(0)}\| \overset{(B.95)}{=} 0$.

Assuming that equation (B.140) is true at the tth iteration, we need to prove

$$\theta^{(t+1)} \leq 4\mu(t+1)\phi^{(t+1)}. \tag{B.141}$$

Applying (B.137) into RHS of (B.141), we have

$$4\mu(t+1)\phi^{(t+1)} \geq \frac{4\mu(t+1)}{1 - \widehat{\delta}'}\left(\phi^{(t)} - \mu\sqrt{(\phi^{(t)})^2 + (\theta^{(t)})^2}\right), \tag{B.142}$$

and substituting (B.138) into LHS of (B.141), we have

$$\theta^{(t+1)} \leq \frac{(4\mu t\phi^{(t)}) + \mu\sqrt{(\phi^{(t)})^2 + (\theta^{(t)})^2}}{1 - \widehat{\delta}'}. \tag{B.143}$$

Then, our goal is to prove RHS of (B.142) is greater than RHS of (B.143). After some manipulations, it is sufficient to show

$$(1 + 4\mu(t+1))\left(\sqrt{(\phi^{(t)})^2 + (\theta^{(t)})^2}\right) \leq 4\phi^{(t)}. \tag{B.144}$$

In the following, we will show that the above relation is true.

First step : We know that

$$4\mu(t+1) \leq 4\mu T \overset{(B.139)}{\leq} 4\eta\rho\mathcal{S}\mathcal{P}(4+62\widehat{c})\widehat{c}\mathcal{T} \overset{(B.87d)(B.139)}{\leq} \frac{4\widehat{c}\eta^2 L_{\max}^2(4+62\widehat{c})}{\kappa\log(\frac{d\kappa}{\delta})} \overset{(a)}{\leq} 1 \tag{B.145}$$

where (a) is true because we choose $c'_{\max} = 1/(2\widehat{c}(4+62\widehat{c}))$ and $\eta \leq c'_{\max}/L_{\max}$.

Second step : Also, we know that

$$4\phi^{(t)} \geq 2\sqrt{2(\phi^{(t)})^2} \overset{(B.140),(B.145)}{\geq} (1+4\mu(t+1))\sqrt{(\phi^{(t)})^2+(\theta^{(t)})^2}.$$

With the above two steps, we have $\theta^{(t+1)} \leq 4\mu(t+1)\phi^{(t+1)}$, which completes the induction.

Recursion of $\phi^{(t)}$:Using (B.140), we have $\theta^{(t)} \overset{(B.140)}{\leq} 4\mu t\phi^{(t)} \overset{(B.145)}{\leq} \phi^{(t)}$, and have

$$(1-\widehat{\delta}')\phi^{(t+1)} \overset{(B.137)}{\geq} \phi^{(t)} - \mu\sqrt{(\phi^{(t)})^2+(\theta^{(t)})^2},$$

which implies

$$\begin{aligned}
\phi^{(t+1)} &\overset{(a)}{\geq} \frac{1}{1-\widehat{\delta}'}\left(\phi^{(t)} - \mu\sqrt{(\phi^{(t)})^2+(\theta^{(t)})^2}\right) \\
&\overset{(b)}{\geq} \frac{1}{1-\frac{\eta\gamma}{2}}\left(\phi^{(t)} - \mu\sqrt{(\phi^{(t)})^2+(\theta^{(t)})^2}\right) \\
&\overset{(c)}{\geq} \frac{1-\frac{\gamma^2\eta^2}{4}}{1-\frac{\eta\gamma}{2}}\phi^{(t)} = (1+\frac{\eta\gamma}{2})\phi^{(t)}
\end{aligned} \tag{B.146}$$

where (a) is true because $1-\widehat{\delta}' > 0$, in (b) we used Corollary 4, i.e., $0 < 1-\widehat{\delta}' \leq 1-\frac{\eta\gamma}{2}$, and (c) is true because $\theta^{(t)} \leq \phi^{(t)}$ and

$$\mu = \eta\rho\mathcal{S}\mathcal{P}(4+62\widehat{c}) \overset{(B.87e)}{\leq} \gamma^2\eta^2\frac{\eta L_{\max}(4+62\widehat{c})}{\log^2(\frac{d\kappa}{\delta})} \overset{(a)}{\leq} \frac{\gamma^2\eta^2}{4\sqrt{2}}$$

where in (a) we choose $c''_{\max} = 1/(4\sqrt{2}(4+62\widehat{c}))$ and $\eta \leq c''_{\max}/L_{\max}$.

Quantifying Escaping Time: From (B.133), we have

$$\begin{aligned}
10\mathcal{S}\widehat{c} \geq \|\mathbf{v}^{(t)}\| \geq \phi^{(t)} &\overset{(B.146)}{\geq} (1+\frac{\gamma\eta}{2})^t\phi^{(0)} \overset{(a)}{\geq} (1+\frac{\gamma\eta}{2})^t\frac{\delta}{2\sqrt{d}}\frac{\eta L_{\max}\mathcal{S}}{\kappa}\log^{-1}(\frac{d\kappa}{\delta}) \\
&\overset{(b)}{\geq} (1+\frac{\gamma\eta}{2})^t\frac{\delta}{2\sqrt{d}}\frac{c\mathcal{S}}{\kappa}\log^{-1}(\frac{d\kappa}{\delta}) \quad \forall t < T
\end{aligned} \tag{B.147}$$

where in (a) we use condition $\upsilon \in [\delta/(2\sqrt{d}), 1]$, in (b) we used $\eta = c/L_{\max}$.

Since (B.147) is true for all $t < T$, we can have

$$\begin{aligned} T - 1 \leq & \frac{\log(20\frac{\widehat{c}}{c}(\frac{\kappa\sqrt{d}}{\delta})\log(\frac{d\kappa}{\delta}))}{\log(1+\frac{\eta\gamma}{2})} \overset{(a)}{<} \frac{4\log(20(\frac{\sqrt{d}d\kappa}{\delta})\frac{\widehat{c}}{c}\log(\frac{d\kappa}{\delta}))}{\eta\gamma} \\ & \overset{(b)}{<} \frac{4\log(20(\frac{d\kappa}{\delta})^2\frac{\widehat{c}}{c})}{\eta\gamma} \overset{(c)}{<} 4(2+\log(20\frac{\widehat{c}}{c}))\mathcal{T} \end{aligned} \tag{B.148}$$

where (a) comes from inequality $\log(1+x) > x/2$ when $x < 1$, in (b) we used relation $\log(x) < x, x > 0$, and (c) is true because $\delta \in (0, \frac{d\kappa}{e}]$ and $\log(d\kappa/\delta) > 1$.

From (B.148), we know that

$$T < 4(2+\log(20\frac{\widehat{c}}{c}))\mathcal{T} + 1 \overset{(a)}{<} 4(2\frac{1}{4}+\log(20\frac{\widehat{c}}{c})\mathcal{T} \tag{B.149}$$

where (a) is true due to the fact that $\eta L_{\max} \geq 1$ and $\log(d\kappa/\delta) > 1$ so we know $\mathcal{T} \geq 1$.

Applying the proof from (B.80) to (B.82), we can also conclude that there exists a universal $\widehat{c}$ such that (B.149) holds. The proof is complete. □

B.3.8 Proof of Lemma 22

First, after the random perturbation, the objective function value in the worst case is increased at most by

$$
\begin{aligned}
& f(\mathbf{u}^{(0)}) - f(\widetilde{\mathbf{x}}^{(t)}) \\
\leq & \sum_{k=1}^{2} \nabla_k f(\widetilde{\mathbf{h}}_{-k}^{(t)}, \widetilde{\mathbf{x}}_k^{(t)})^T \xi_k + \frac{L_k}{2}\|\xi_k\|^2 \\
\leq & \sum_{k=1}^{2} \left(\nabla_k f(\widetilde{\mathbf{h}}_{-k}^{(t)}, \widetilde{\mathbf{x}}_k^{(t)}) - \nabla_k f(\widetilde{\mathbf{h}}_{-k}^{(t)}, \widetilde{\mathbf{x}}_k^{(t+1)})\right)^T \xi_k + \sum_{k=1}^{2} \nabla_k f(\widetilde{\mathbf{h}}_{-k}^{(t)}, \widetilde{\mathbf{x}}_k^{(t+1)})^T \xi_k + \frac{L_k}{2}\|\xi_k\|^2 \\
\leq & \sum_{k=1}^{2} L_{\max} \left\|\mathbf{x}_k^{(t+1)} - \mathbf{x}_k^{(t)}\right\| \|\xi_k\| + \sum_{k=1}^{2} \|\nabla_k f(\widetilde{\mathbf{h}}_{-k}^{(t)}, \widetilde{\mathbf{x}}_k^{(t+1)})\| \|\xi_k\| + \frac{L_{\max}}{2}\|\xi\|^2 \\
\overset{(a)}{\leq} & 1.25 \sum_{k=1}^{2} \|\nabla_k f(\widetilde{\mathbf{h}}_{-k}^{(t)}, \widetilde{\mathbf{x}}_k^{(t+1)})\| \|\xi_k\| + \frac{L_{\max}}{2}\|\xi\|^2 \\
\overset{(b)}{\leq} & 1.25\|\xi\| \sqrt{\sum_{k=1}^{2} 2\|\nabla_k f(\widetilde{\mathbf{h}}_{-k}^{(t)}, \widetilde{\mathbf{x}}_k^{(t+1)})\|^2} + \frac{L_{\max}}{2}\|\xi\|^2 \\
\overset{(c)}{\leq} & 1.25 \frac{\mathcal{G}}{\kappa} \frac{\eta L_{\max} \mathcal{S}}{\kappa \log(\frac{d\kappa}{\delta})\mathcal{P}} + \frac{L_{\max}}{2} (\frac{\eta L_{\max} \mathcal{S}}{\kappa \log(\frac{d\kappa}{\delta})\mathcal{P}})^2 \leq \frac{3}{2}\mathcal{F}
\end{aligned}
\tag{B.150}
$$

where $\mathbf{u}^{(0)}$ is a vector that follows uniform distribution within the ball $\mathbb{B}^{(d)}_{\widetilde{\mathbf{x}}^{(t)}}(r)$, $\mathbb{B}^{(d)}_{\widetilde{\mathbf{x}}^{(t)}}$ denotes the d-dimensional ball centered at $\widetilde{\mathbf{x}}^{(t)}$ with radius r, ξ_k represents the kth block of the vector which is the difference between random generated vector $\mathbf{u}^{(0)}$ and saddle point $\widetilde{\mathbf{x}}^{(t)}$, and in (a) we choose $\eta \leq 1/(4L_{\max})$ and (b) is true because $\xi := [\xi_1, \ldots, \xi_K]$, $\|\xi_k\| \leq \|\xi\|, \forall k$, and in (c) we used $\kappa > 1$, $\log(d\kappa/\delta) > 1$, $\mathcal{P} \geq 2$ and Condition 2 where g_{th} is defined in (B.99).

Then, the rest of proof of Lemma 22 is the same as the rest of Lemma 16, therefore ignored for simplicity.

B.4 Proof of Lemma 7

Proof. Consider function

$$
f(\mathbf{x}) = \mathbf{x}^T \mathbf{A} \mathbf{x} + \frac{1}{4}\|\mathbf{x}\|_4^4 \tag{B.151}
$$

where $\mathbf{x} \in \mathcal{S}$, $\mathcal{S} = \{\mathbf{x} | \|\mathbf{x}\|^2 \leq \tau\}$ and $\tau \geq \lambda_{\max}(\mathbf{A})$.

To prove L-smooth Lipschitz continuity :

$$\begin{aligned}\|\nabla f(\mathbf{x})-\nabla f(\mathbf{y})\| =& \left\| 2(\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{y}) + \begin{bmatrix} x_1^3-y_1^3 \\ \vdots \\ x_d^3-y_d^3 \end{bmatrix} \right\|, \quad \forall \mathbf{x},\mathbf{y}\in\mathcal{S} \\ \leq & 2\lambda_{\max}(\mathbf{A})\|\mathbf{x}-\mathbf{y}\| + \left\| \begin{bmatrix} (x_1-y_1)(x_1^2+x_1y_1+y_1^2) \\ \vdots \\ (x_d-y_d)(x_d^2+x_dy_d+y_d^2) \end{bmatrix} \right\| \\ \overset{(a)}{\leq} & 2\tau\|\mathbf{x}-\mathbf{y}\| + 3\tau\|\mathbf{x}-\mathbf{y}\| \leq 5\tau\|\mathbf{x}-\mathbf{y}\|\end{aligned}$$

where x_i denotes the ith entry of vector $\mathbf{x}$, and (a) is true because

$$x_i^2\leq\tau, \quad y_i^2\leq\tau, \quad x_iy_i\leq(x_i^2+y_i^2)/2\leq\tau, \forall i. \tag{B.152}$$

To prove block-wise Lipschitz continuity : Without loss of generality, consider first block $\mathbf{x}_1\in\mathcal{S}'$ where $\mathcal{S}'=\{\mathbf{x}_1|\|\mathbf{x}_1\|^2\leq\tau',\mathbf{x}_1\in\mathbb{R}^{d'\times 1}\}$ and d' denotes the dimension of $\mathbf{x}_1$. Consider $\tau'\geq\lambda_{\max}(\mathbf{A}')$ where $\mathbf{A}'\in\mathbb{R}^{d'\times d'}$ is the leading principal minor of matrix $\mathbf{A}$ of order d'. Obviously, we have $\tau'\leq\tau$.

$$\begin{aligned}&\|\nabla_1 f(\mathbf{x}_{-1},\mathbf{x}_1)-\nabla_1 f(\mathbf{x}_{-1},\mathbf{x}_1')\| \\ =& \left\| 2\mathbf{I}_1'\left(\mathbf{A}\begin{bmatrix}\mathbf{x}_1\\ \mathbf{x}_{-1}\end{bmatrix} - \mathbf{A}\begin{bmatrix}\mathbf{x}_1'\\ \mathbf{x}_{-1}'\end{bmatrix}\right) + \begin{bmatrix} x_1^3-x_1'^3 \\ \vdots \\ x_{d'}^3-x_{d'}'^3 \end{bmatrix} \right\|, \quad \forall \mathbf{x},\mathbf{x}'\in\mathcal{S}' \\ \leq & 2\|\mathbf{I}_1'\left(\mathbf{A}\begin{bmatrix}\mathbf{x}_1\\ \mathbf{x}_{-1}\end{bmatrix} - \mathbf{A}\begin{bmatrix}\mathbf{x}_1'\\ \mathbf{x}_{-1}\end{bmatrix}\right)\| + \left\| \begin{bmatrix} (x_1-x_1')(x_1^2+x_1x_1'+x_1'^2) \\ \vdots \\ (x_{d'}-x_{d'}')(x_{d'}^2+x_{d'}x_{d'}'+x_{d'}'^2) \end{bmatrix} \right\| \\ \overset{(a)}{\leq} & 2\lambda_{\max}(\mathbf{A}')\|\mathbf{x}_1-\mathbf{x}_1'\| + 3\tau'\|\mathbf{x}_1-\mathbf{y}_1\| \\ \leq & 5\tau'\|\mathbf{x}_1-\mathbf{x}_1'\|, \quad \forall \mathbf{x},\mathbf{x}'\end{aligned}$$

where (a) is true because we used $\mathbf{I}_1' := \begin{bmatrix} \mathbf{I}_{d'} & 0 \\ 0 & 0 \end{bmatrix}$ which selects the first d' rows of $\mathbf{A}([\mathbf{x}_1; \mathbf{x}_{-1}] - [\mathbf{x}_1'; \mathbf{x}_{-1}])$.

To prove Hessian Lipschitz continuity :

$$\begin{aligned} \|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y})\| =& 3 \left\| \begin{array}{ccc} x_1^2 - y_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & x_d^2 - y_d^2 \end{array} \right\| \\ \leq& 6\sqrt{\tau} \left\| \begin{array}{ccc} x_1 - y_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & x_d - y_d \end{array} \right\| = 6\sqrt{\tau}\|\mathbf{x} - \mathbf{y}\| \end{aligned}$$

where (a) is true because $x_i + y_i \leq \sqrt{(x_i + y_i)^2} = \sqrt{x_1^2 + 2x_i y_i + y_i^2} \overset{(B.152)}{\leq} 2\sqrt{\tau}, \forall i.$ □